5 Bayesian Statistical Inference

"The Bayesian approach is the numerical realization of common sense." (Bayesians)

We have already addressed the main philosophical differences between classical and Bayesian statistical inferences in §4.1. In this chapter, we introduce the most important aspects of Bayesian statistical inference and techniques for performing such calculations in practice. We first review the basic steps in Bayesian inference in §5.1–5.4, and then illustrate them with several examples in §5.6–5.7. Numerical techniques for solving complex problems are discussed in §5.8, and the last section provides a summary of pros and cons for classical and Bayesian methods.

Let us briefly note a few historical facts. The Reverend Thomas Bayes (1702–1761) was a British amateur mathematician who wrote a manuscript on how to combine an initial belief with new data to arrive at an improved belief. The manuscript was published posthumously in 1763 and gave rise to the name Bayesian statistics. However, the first renowned mathematician to popularize Bayesian methodology was Pierre Simon Laplace, who rediscovered (1774) and greatly clarified Bayes' principle. He applied the principle to a variety of contemporary problems in astronomy, physics, population statistics, and even jurisprudence. One of the most famous results is his estimate of the mass of Saturn and its uncertainty, which remain consistent with the best measurements of today.

Despite Laplace's fame, Bayesian analysis did not secure a permanent place in science. Instead, classical frequentist statistics was adopted as the norm (this could be at least in part due to the practical difficulties of performing full Bayesian calculations without the aid of computers). Much of Laplace's Bayesian analysis was ignored until the early twentieth century when Harold Jeffreys reinterpreted Laplace's work with much clarity. Yet, even Jeffreys' work was not fully comprehended until around 1960, when it took off thanks to vocal proponents such as de Finetti, Savage, Wald, and Jaynes, and of course, the advent of computing technology. Today, a vast amount of literature exists on various Bayesian topics, including the two books by Jaynes and Gregory listed in §1.3. For a very informative popular book about the resurgence of Bayesian methods, see [26].

5.1. Introduction to the Bayesian Method

The basic premise of the Bayesian method is that probability statements are not limited to data, but can be made for model parameters and models themselves. Inferences are made by producing probability density functions (pdfs); most notably, model parameters are treated as random variables.

The Bayesian method has gained wide acceptance over the last few decades, in part due to maturing development of its philosophical and technical foundations, and in part due to the ability to actually perform the required computations. The Bayesian method yields optimal results, given all the available and explicitly declared information, assuming, of course, that all of the supplied information is correct. Even so, it is not without its own pitfalls, as discussed at the end of this chapter.

Classical and Bayesian techniques share an important ingredient: the data likelihood function introduced in §4.2. In classical statistics, the data likelihood function is used to find model parameters that yield the highest data likelihood. Yet, the likelihood function cannot be interpreted as a probability density function for model parameters. Indeed, the pdf for a model parameter is not even a valid concept in classical statistics. The Bayesian method extends the concept of the data likelihood function by adding extra, so-called *prior*, information to the analysis, and assigning pdfs to all model parameters and models themselves.

We use the following simple example for motivating the inclusion of prior information (problem 48 in Lup93). In this example, the maximum likelihood approach might be thought to give us an unsatisfying answer, whereas the addition of "side" or prior information can improve the inference.

Imagine you arrive at a bus stop, and observe that the bus arrives t minutes later (it is assumed that you had no knowledge about the bus schedule). What is the mean time between two successive buses, τ, if the buses keep a regular schedule? It is easy to derive an intuitive answer. The wait time is distributed uniformly in the interval $0 \leq t \leq \tau$, and on average you would wait for $t = \tau/2$ minutes. Rearranging this gives $\tau = 2t$, which agrees with intuition.

What does the maximum likelihood approach give? The probability that you will wait t minutes (the likelihood of data) is given by the uniform distribution

$$p(t|\tau) = \begin{cases} 1/\tau & \text{if } 0 \leq t \leq \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

Because we only observe a single point, the data likelihood (eq. 4.1) is simply equal to this probability. The maximum likelihood, then, corresponds to the smallest possible τ such that $t \leq \tau$: this is satisfied by $\tau = t$ and not $\tau = 2t$ as we expected! Computing the expectation value or the median for τ does not help either because the resulting integrals diverge. These puzzling results are resolved by the use of appropriate prior information, as discussed in the next section. We shall see several other examples where the addition of extra information changes the results we would get from the maximum likelihood approach.

The Bayesian method is not, however, motivated by the differences in results between maximum likelihood and Bayesian techniques. These differences are often negligible, especially when the data sets are large. Rather, the Bayesian method is

motivated by its ability to provide a full probabilistic framework for data analysis. One of the most important aspects of Bayesian analysis is the ability to straightforwardly incorporate unknown or uninteresting model parameters—the so-called *nuisance parameters*—in data analysis. But let us start with basics and introduce the Bayesian framework step by step. After reviewing the basic analysis steps, we shall illustrate them with practical examples collected in §5.6–5.7.

5.1.1. The Essence of the Bayesian Idea

We already introduced Bayes' rule (eqs. 3.10 and 3.11) in §3.1.3. Bayes' rule is simply a mathematical identity following from a straightforward application of the rules of probability and thus is not controversial in and of itself. The frequentist vs. Bayesian controversy sets in when we apply Bayes' rule to the likelihood function $p(D|M)$ to obtain Bayes' theorem:

$$p(M|D) = \frac{p(D|M)\, p(M)}{p(D)}, \tag{5.2}$$

where D stands for data, and M stands for model. Bayes' theorem quantifies the rule for "combining an initial belief with new data to arrive at an improved belief" and says that "improved belief" is proportional to the product of "initial belief" and the probability that "initial belief" generated the observed data.

To be more precise, let us explicitly acknowledge the presence of prior information I and the fact that models are typically described by parameters whose values we want to estimate from data:

$$p(M, \boldsymbol{\theta}|D, I) = \frac{p(D|M, \boldsymbol{\theta}, I)\, p(M, \boldsymbol{\theta}|I)}{p(D|I)}. \tag{5.3}$$

In general, as we saw in the likelihood function in §4.2.1, the model M includes k model parameters θ_p, $p = 1, \ldots, k$, abbreviated as vector $\boldsymbol{\theta}$ with components θ_p. Strictly speaking, the vector $\boldsymbol{\theta}$ should be labeled by M since different models may be described by different parameters. We will sometimes write just M or $\boldsymbol{\theta}$ (θ in one-dimensional cases), depending on what we are trying to emphasize.

The result $p(M, \boldsymbol{\theta}|D, I)$ is called the *posterior* pdf for model M *and* parameters $\boldsymbol{\theta}$, given data D and other prior information I. This term is a $(k + 1)$-dimensional pdf in the space spanned by k model parameters and the model index M. The term $p(D|M, \boldsymbol{\theta}, I)$ is, as before, the *likelihood* of data *given* some model M and given some fixed values of parameters $\boldsymbol{\theta}$ describing it, and all other prior information I. The term $p(M, \boldsymbol{\theta}|I)$ is the a priori joint probability for model M and its parameters $\boldsymbol{\theta}$ in the absence of any of the data used to compute likelihood, and is often simply called the *prior*. The term prior is understood logically and not temporally: despite its name, measurements D that enter into the calculation of data likelihood may be collected before information that is used to construct prior $p(M, \boldsymbol{\theta}|I)$. The prior can be expanded as

$$p(M, \boldsymbol{\theta}|I) = p(\boldsymbol{\theta}|M, I)\, p(M|I), \tag{5.4}$$

and in parameter estimation problems we need only specify $p(\boldsymbol{\theta}|M, I)$. In the context of the model selection, however, we need the full prior.

The term $p(D|I)$ is the *probability of data*, or the prior predictive probability for D. It provides proper normalization for the posterior pdf and usually it is not explicitly computed when estimating model parameters: rather, $p(M, \boldsymbol{\theta}|D, I)$ for a given M is simply renormalized so that its integral over all model parameters $\boldsymbol{\theta}$ is unity. The integral of the prior $p(\boldsymbol{\theta}|M, I)$ over all parameters should also be unity, but for the same reason, calculations of the posterior pdf are often done with an arbitrary normalization. An important exception is model selection discussed below, where the correct normalization of the product $p(D|M, \boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}|M, I)$ is crucial.

5.1.2. Discussion

Why is this interpretation controversial? If we want to write "$p(M, \boldsymbol{\theta}|D, I)$," we must somehow acknowledge that it is not a probability in the same sense as $p(D|M, \boldsymbol{\theta}, I)$. The latter can be easily understood in terms of the long-term frequency of events, for example as the long-term probability of observing heads a certain number of times given a certain physical coin (hence the term "frequentist"). If we measure the mass of a planet, or an apple, or an elementary particle, this mass is not a random number; it is what it is and cannot have a distribution! To acknowledge this, we accept that when using the Bayesian formalism, $p(M, \boldsymbol{\theta}|D, I)$ corresponds to the state of our *knowledge* (i.e., belief) about a model and its parameters, given data D and prior information I. This change of interpretation of the symbols (note that the mathematical axioms of probability do not change under Bayesianism) introduces the notion of the posterior probability distribution for models and model parameters (as Laplace did for the mass of Saturn in one of the first applications of the Bayesian method[1]).

Let us return to our bus stop example and see how the prior helps. Eq. 5.1 corresponds to $p(D|M, \boldsymbol{\theta}, I)$ in eq. 5.2, with the vector of parameters $\boldsymbol{\theta}$ having only one component: τ. We take the prior $p(\tau|I)$ as proportional to τ^{-1} for reasons that will be explained in §5.2.1 (τ here is a scale parameter). The posterior probability density function for τ then becomes

$$p(\tau|t, I) = \begin{cases} t/\tau^2 & \text{if } \tau \geq t, \\ 0 & \text{otherwise.} \end{cases} \tag{5.5}$$

This posterior pdf is a result of multiplying the likelihood and prior, which are both proportional to τ^{-1}, and then normalizing the integral of $p(\tau|t, I)$ as

$$\int_t^\infty p(\tau|t, I)\, d\tau = \int_t^\infty \frac{C}{\tau^2} d\tau = 1, \tag{5.6}$$

which gives $C = t$ as the normalization constant. The divergent integral over τ encountered in likelihood analysis is mitigated by the extra τ^{-1} term from the prior. The median τ given by the posterior $p(\tau|t, I)$ is now equal to $2t$, in agreement with our expectations. An interesting side result is that the p% quantiles are

[1] Laplace said: "I find that it is a bet of 11,000 against one that the error of this result is not 1/100th of its value ..." (see p. 79 in [25]). Therefore, Laplace clearly interpreted measurements as giving a probability statement about the mass of Saturn, although there is only one Saturn and its true mass is what it is, and it is not a random variable according to classical statistics.

equal to $(1 - t/\tau_p)$; for example, the 95% confidence region for τ, known as the *credible region* in the Bayesian context, spans $1.03\,t < \tau < 40\,t$. If we waited for a bus for 1 minute, then, adopting the usual 95% confidence region, we cannot reject the possibility that τ is as large as 40 minutes. Equivalently, if we waited for a bus for 1 minute, we can paraphrase Laplace and say that "it is a bet of 20 against 1 that the bus will arrive in the interval between 0.03 minutes and 39 minutes from now."

We will now discuss the main ingredients of the Bayesian methodology in more detail, and then illustrate the main concepts using a few simple practical examples in §5.6–5.7.

5.1.3. Summary of the Bayesian Statistical Inference

To simplify the notation, we will shorten $M(\boldsymbol{\theta})$ to simply M whenever the absence of explicit dependence on $\boldsymbol{\theta}$ is not confusing. A completely Bayesian data analysis has the following conceptual steps:

1. The formulation of the data likelihood $p(D|M, I)$. Of course, if the adopted $p(D|M, I)$ is a poor description of this process, then the resulting posterior pdf will be inaccurate, too.
2. The choice of the prior $p(\boldsymbol{\theta}|M, I)$, which incorporates all other knowledge that might exist, but is *not* used when computing the likelihood (e.g., prior measurements of the same type, different measurements, or simply an uninformative prior, as discussed below).
3. Determination of the posterior pdf, $p(M|D, I)$, using Bayes' theorem. In practice, this step can be computationally intensive for complex multidimensional problems. Often $p(D|I)$ is not explicitly specified because $p(M|D, I)$ can be properly normalized by renormalizing the product $p(D|M, I)\, p(M|I)$.
4. The search for the best model parameters M, which maximize $p(M|D, I)$, yielding the *maximum a posteriori* (MAP) estimate. This *point estimate* is the natural analog to the maximum likelihood estimate (MLE) from classical statistics. Another natural Bayesian estimator is the *posterior mean*:

$$\overline{\theta} = \int \theta\; p(\theta|D)\, d\theta. \tag{5.7}$$

   In multidimensional cases, $p(\theta|D)$, where θ is one of many model parameters, is obtained from $p(M, \boldsymbol{\theta}|D, I)$ using *marginalization*, or integration of $p(M, \boldsymbol{\theta}|D, I)$ over all other model parameters and renormalization ($\int p(\theta|D)\, d\theta = 1$, and the model index is not explicitly acknowledged because it is implied by the context). Both the MAP and posterior mean are only convenient ways to summarize the information provided by the posterior pdf, and often do not capture its full information content.
5. Quantification of uncertainty in parameter estimates, via *credible regions* (the Bayesian counterpart to frequentist confidence regions). As in MLE, such an estimate can be obtained analytically by doing mathematical derivations specific to the chosen model. Also as in MLE, various numerical techniques can be used to simulate samples from the posterior. This can be viewed as an analogy to the frequentist bootstrap approach, which can simulate

draws of samples from the true underlying distribution of the data. In both cases, various descriptive statistics can then be computed on such samples to examine the uncertainties surrounding the data and estimators of model parameters based on that data.

6. *Hypothesis testing* as needed to make other conclusions about the model or parameter estimates. Unlike hypothesis tests in classical statistics, in Bayesian inference hypothesis tests incorporate the prior and thus may give different results.

The Bayesian approach can be thought of as formalizing the process of continually refining our state of knowledge about the world, beginning with no data (as encoded by the prior), then updating that by multiplying in the likelihood once the data D are observed to obtain the posterior. When more data are taken, then the posterior based on the first data set can be used as the prior for the second analysis. Indeed, the data sets can be fundamentally different: for example, when estimating cosmological parameters using observations of supernovas, the prior often comes from measurements of the cosmic microwave background, the distribution of large-scale structure, or both (e.g., [18]). This procedure is acceptable as long as the pdfs refer to the same quantity. For a pedagogical discussion of probability calculus in a Bayesian context, please see [14].

5.2. Bayesian Priors

How do we choose the prior[2] $p(\boldsymbol{\theta}|I) \equiv p(\boldsymbol{\theta}|M, I)$ in eq. 5.4? The prior incorporates all other knowledge that might exist, but is not used when computing the likelihood, $p(D|M, \boldsymbol{\theta}, I)$. To reiterate, despite the name, the data may chronologically precede the information in the prior. The latter can include the knowledge extracted from prior measurements of the same type as the data at hand, or different measurements that constrain the same quantity whose posterior pdf we are trying to constrain with the new data. For example, we may know from older work that the mass of an elementary particle is m_A, with a Gaussian uncertainty parametrized by σ_A, and now we wish to utilize a new measuring apparatus or method. Hence, m_A and σ_A may represent a convenient summary of the posterior pdf from older work that is now used as a prior for the new measurements. Therefore, the terms prior and posterior do not have an absolute meaning. Such priors that incorporate information based on other measurements (or other sources of meaningful information) are called *informative priors*.

5.2.1. Priors Assigned by Formal Rules

When no other information, except for the data we are analyzing, is available, we can assign priors by formal rules. Sometimes these priors are called *uninformative priors* but this term is a misnomer because these priors can incorporate weak but objective information such as "the model parameter describing variance cannot be negative."

[2] While common in the physics literature, the adjective "Bayesian" in front of "prior" is rare in the statistics literature.

Note that even the most uninformative priors still affect the estimates, and the results are not generally equivalent to the frequentist or maximum likelihood estimates.

As an example, consider a *flat prior*,

$$p(\theta|I) \propto C, \tag{5.8}$$

where $C > 0$ is a constant. Since $\int p(\theta|I)\, d\theta = \infty$, this is not a pdf; this is an example of an *improper prior*. In general, improper priors are not a problem as long as the resulting posterior is a well-defined pdf (because the likelihood effectively controls the result of integration). Alternatively, we can adopt a lower and an upper limit on θ which will prevent the integral from diverging (e.g., it is reasonable to assume that the mass of a newly discovered elementary particle must be positive and smaller than the Earth's mass). Flat priors are sometimes considered ill defined because a flat prior on a parameter does not imply a flat prior on a transformed version of the parameter (e.g., if $p(\theta)$ is a flat prior, $\ln\theta$ does not have a flat prior).

Although uninformative priors do not contain specific information, they can be assigned according to several general principles. The main point here is that for the same prior information, these principles result in assignments of the same priors.

The oldest method is the *principle of indifference* which states that a set of basic, mutually exclusive possibilities need to be assigned equal probabilities (e.g., for a fair six-sided die, each of the outcomes has a prior probability of 1/6). The *principle of consistency*, based on transformation groups, demands that the prior for a location parameter should not change with translations of the coordinate system, and yields a flat prior. Similarly, the prior for a scale parameter should not depend on the choice of units. If the scale parameter is σ and we rescale our measurement units by a positive factor a, we get a constraint

$$p(\sigma|I)\, d\sigma = p(a\sigma|I)\, d(a\sigma). \tag{5.9}$$

The solution is $p(\sigma|I) \propto \sigma^{-1}$ (or a flat prior for $\ln\sigma$), called a scale-invariant prior.

When we have additional weak prior information about some parameter, such as a low-order statistic, we can use the *principle of maximum entropy* to construct priors consistent with that information.

5.2.2. The Principle of Maximum Entropy

Entropy measures the information content of a pdf. We shall use S as the symbol for entropy, although we have already used s for the sample standard deviation (eq. 3.32), because we never use both in the same context. Given a pdf defined by N discrete values p_i, with $\sum_{i=1}^{N} p_i = 1$, its entropy is defined as

$$S = -\sum_{i=1}^{N} p_i \ln(p_i) \tag{5.10}$$

(note that $\lim_{p\to 0} [p \ln p] = 0$). This particular functional form can be justified using arguments of logical consistency (see Siv06 for an illuminating introduction) and information theory (using the concept of minimum description length, see HTF09). It is also called Shannon's entropy because Shannon was the first one to derive it

in the context of information in 1948. It resembles thermodynamic entropy: this observation is how it got its name (this similarity is not coincidental; see Jay03). The unit for entropy is the *nat* (from natural unit; when ln is replaced by the base 2 logarithm, then the unit is the more familiar *bit*; 1 nat = 1.44 bits).

Sivia (see Siv06) discusses the derivation of eq. 5.10 and its extension to the continuous case

$$S = -\int_{-\infty}^{\infty} p(x) \ln\left(\frac{p(x)}{m(x)}\right) dx, \tag{5.11}$$

where the "measure" $m(x)$ ensures that entropy is invariant under a change of variables.

The idea behind the principle of maximum entropy for assigning uninformative priors is that by maximizing the entropy over a suitable set of pdfs, we find the distribution that is least informative (given the constraints). The power of the principle comes from a straightforward ability to add additional information about the prior distribution, such as the mean value and variance. Computational details are well exposed in Siv06 and Greg05, and here we only review the main results.

Let us start with Sivia's example of a six-faced die, where we need to assign six prior probabilities. When no specific information is available, the principle of indifference states that each of the outcomes has a prior probability of 1/6. If additional information is available (with its source unspecified), such as the mean value of a large number of rolls, μ, (for a fair die the expected mean value is 3.5), then we need to adjust prior probabilities to be consistent with this information. Given the six probabilities p_i, the expected mean value is

$$\sum_{i=1}^{6} i\ p_i = \mu, \tag{5.12}$$

and of course

$$\sum_{i=1}^{6} p_i = 1. \tag{5.13}$$

We have two constraints for the six unknown values p_i. The problem of assigning individual p_i can be solved using the principle of maximum entropy and the method of Lagrangian multipliers. We need to maximize the following quantity with respect to six individual p_i:

$$Q = S + \lambda_0 \left(1 - \sum_{i=1}^{6} p_i\right) + \lambda_1 \left(\mu - \sum_{i=1}^{6} i p_i\right), \tag{5.14}$$

where the first term is entropy:

$$S = -\sum_{i=1}^{6} p_i \ln\left(\frac{p_i}{m_i}\right), \tag{5.15}$$

and the second and third term come from additional constraints (λ_0 and λ_1 are called Lagrangian multipliers). In the expression for entropy, m_i are the values that would be assigned to p_i in the case when no additional information is known (i.e., without constraint on the mean value; in this problem $m_i = 1/6$). By differentiating Q with respect to p_i, we get conditions

$$-\left[\ln\left(\frac{p_i}{m_i}\right)+1\right]-\lambda_0 - i\lambda_1 = 0, \tag{5.16}$$

and solutions

$$p_i = m_i \exp(-1-\lambda_0)\exp(i\lambda_1). \tag{5.17}$$

The two remaining unknown values of λ_0 and λ_1 can be determined numerically using constraints given by eqs. 5.12 and 5.13. Therefore, *although our knowledge about p_i is incomplete and based on only two constraints, we can assign all six p_i!* When the number of possible discrete events is infinite (as opposed to six here), the maximum entropy solution for assigning p_i is the Poisson distribution parametrized by the expectation value μ.

In the corresponding continuous case, the maximum entropy solution for the prior is

$$p(\theta|\mu) = \frac{1}{\mu}\exp\left(\frac{-\theta}{\mu}\right). \tag{5.18}$$

This result is based on the constraint that we only know the expectation value for θ ($\mu = \int \theta p(\theta)\, d\theta$), and assuming a flat distribution $m(\theta)$ (the prior for θ when the additional constraint given by μ is not imposed). Another useful result is that when only the mean and the variance are known in advance, with the distribution defined over the whole real line, the maximum entropy solution is a Gaussian distribution with those values of mean and variance.

A quantity closely related to entropy is the Kullback–Leibler (KL) divergence from $p(x)$ to $m(x)$,

$$\mathrm{KL} = \sum_i p_i \ln\left(\frac{p_i}{m_i}\right), \tag{5.19}$$

and analogously for the continuous case (i.e., KL is equal to S from eq. 5.11 except for the minus sign).

Sometimes, the KL divergence is called the KL distance between two pdfs. However, the KL distance is not a true distance metric because its value is not the same when $p(x)$ and $m(x)$ are switched. In Bayesian statistics the KL divergence can be used to measure the information gain when moving from a prior distribution to a posterior distribution. In information theory, the KL divergence can be interpreted as the additional message-length per datum if the code that is optimal for $m(x)$ is used to transmit information about $p(x)$. The KL distance will be discussed in a later chapter (see §9.7.1).

5.2.3. Conjugate Priors

In special combinations of priors and likelihood functions, the posterior probability has the same functional form as the prior probability. These priors are called *conjugate priors* and represent a convenient way for generalizing computations.

When the likelihood function is a Gaussian, then the conjugate prior is also a Gaussian. If the prior is parametrized as $\mathcal{N}(\mu_p, \sigma_p)$, and the data can be summarized as $\mathcal{N}(\overline{x}, s)$ (see eqs. 3.31 and 3.32), then the posterior[3] is $\mathcal{N}(\mu^0, \sigma^0)$, with

$$\mu^0 = \frac{\mu_p/\sigma_p^2 + \overline{x}/s^2}{1/\sigma_p^2 + 1/s^2} \quad \text{and} \quad \sigma^0 = \left(1/\sigma_p^2 + 1/s^2\right)^{-1/2}. \tag{5.20}$$

If the data have a smaller scatter (s) than the width of the prior (σ_p), then the resulting posterior (i.e., μ^0) is closer to $\overline{x}$ than to μ_p. Since μ^0 is obviously *different* from $\overline{x}$, this Bayesian estimator is biased! On the other hand, if we choose a very informative prior with $\sigma_p \ll s$, then the data will have little impact on the resulting posterior and μ^0 will be much closer to μ_p than to $\overline{x}$.

In the discrete case, the most frequently encountered conjugate priors are the beta distribution for binomial likelihood, and the gamma distribution for Poissonian likelihood (refer to §3.3 for descriptions of these distributions). For a more detailed discussion, see Greg05. We limit discussion here to the first example.

The beta distribution (see §3.3.10) allows for more flexibility when additional information about discrete measurements, such as the results of prior measurements, is available. A flat prior corresponds to $\alpha = 1$ and $\beta = 1$. When the likelihood function is based on a binomial distribution described by parameters N and k (see §3.3.3), and the prior is the beta distribution, then the posterior is also a beta distribution. It can be shown that parameters describing the posterior are given by $\alpha^0 = \alpha_p + k$ and $\beta^0 = \beta_p + N - k$, where α_p and β_p describe the prior. Evidently, as both k and $N - k$ become much larger than α_p and β_p, the "memory" of the prior information is, by and large, gone. This behavior is analogous to the case with $s \ll \sigma_p$ for the Gaussian conjugate prior discussed above.

5.2.4. Empirical and Hierarchical Bayes Methods

Empirical Bayes refers to an approximation of the Bayesian inference procedure where the parameters of priors (or *hyperparameters*) are estimated from the data. It differs from the standard Bayesian approach, in which the parameters of priors are chosen before any data are observed. Rather than integrate out the hyperparameters as in the standard approach, they are set to their most likely values. Empirical Bayes is also sometimes known as *maximum marginal likelihood*; for more details, see [1].

The empirical Bayes method represents an approximation to a fully Bayesian treatment of a *hierarchical Bayes* model. In hierarchical, or multilevel, Bayesian analysis a prior distribution depends on unknown variables, the hyperparameters, that describe the group (population) level probabilistic model. Their priors, called

[3]The posterior pdf is by definition normalized to 1. However, the product of two Gaussian functions, before renormalization, has an extra multiplicative term compared to $\mathcal{N}(\mu^0, \sigma^0)$. Strictly speaking, the product of two Gaussian pdfs is not a Gaussian pdf because it is not properly normalized.

hyperpriors, resemble the priors in simple (single-level) Bayesian models. This approach is useful for quantifying uncertainty in various aspects of the pdf for the prior, and for using overall population properties when estimating parameters of a single population member (for a detailed discussion, see [9]). Hierarchical Bayes models are especially useful for complex data analysis problems because they can effectively handle multiple sources of uncertainty at all stages of the data anaysis. For a recent application of hierarchical Bayes modeling in astronomy, see [17].

5.3. Bayesian Parameter Uncertainty Quantification

The posterior pdf in the Bayesian framework is treated as any other probabilistic pdf. In practice, it is often summarized in terms of various point estimates (e.g., MAP) or in terms of parameter intervals defined by certain properties of cumulative probability. Although very useful in practice, these summaries rarely capture the full information content of a posterior pdf.

5.3.1. Posterior Intervals

To obtain a Bayesian *credible region* estimate, we find a and b such that $\int_{-\infty}^{a} f(\theta)\,d\theta = \int_{b}^{\infty} f(\theta)\,d\theta = \alpha/2$. Then the probability that the true value of parameter θ is in the interval (a, b) is equal to $1 - \alpha$, in analogy with the classical confidence intervals, and the interval (a, b) is called a $1 - \alpha$ *posterior interval*.

In practice, the posterior pdf, $p(\theta)$, is often not an analytic function (e.g., it can only be evaluated numerically). In such cases, we can compute statistics such as the posterior mean and the $1 - \alpha$ posterior interval, using *simulation* (sampling). If we know how to compute $p(\theta)$, then we can use the techniques for random number generation described in §3.7 to draw N values θ_j. Then we can approximate the posterior mean as the sample mean, and approximate the $1 - \alpha$ posterior interval by finding $\alpha/2$ and $(1-\alpha/2)$ sample quantiles. Several examples of this type of procedure in one and two dimensions are found in §5.6.

5.3.2. Marginalization of Parameters

Consider a problem where the model in the posterior pdf, $p(M, \boldsymbol{\theta}|D, I)$, is parametrized by a vector of k free parameters, $\boldsymbol{\theta}$. A subset of these parameters are of direct interest, while the remaining parameters are used to describe certain aspects of data collection that are not of primary interest. For example, we might be measuring the properties of a spectral line (position, width, strength) detected in the presence of an unknown and variable background. We need to account for the background when describing the statistical behavior of our data, but the main quantities of interest are the spectral line properties. In order to obtain the posterior pdf for each interesting parameter, we can integrate the multidimensional posterior pdf over all other parameters. Alternatively, if we want to understand covariances between interesting parameters, we can integrate the posterior pdf only over uninteresting parameters called *nuisance parameters*. This integration procedure is known as *marginalization* and the resulting pdf is called the *marginal posterior pdf*. An analog of marginalization of parameters does not exist in classical statistics.

We have already discussed multidimensional pdfs and marginalization in the context of conditional probability (in §3.1.3). An example of integrating a two-dimensional pdf to obtain one-dimensional marginal distributions is shown in figure 3.2. Let us assume that x in that figure corresponds to an interesting parameter, and y is a nuisance parameter. The right panels show the posterior pdfs for x if somehow we knew the value of the nuisance parameter, for three different values of the latter. When we do not know the value of the nuisance parameter, we integrate over all plausible values and obtain the marginalized posterior pdf for x, shown at the bottom of the left panel. Note that the marginalized pdf spans a wider range of x than the three pdfs in the right panel. This difference is a general result.

Several practical examples of Bayesian analysis discussed in §5.6 use and illustrate the concept of marginalization.

5.4. Bayesian Model Selection

Bayes' theorem as introduced by eq. 5.3 quantifies the posterior pdf of parameters describing a single model, with that model *assumed to be true*. In model selection and hypothesis testing, we formulate alternative scenarios and ask which ones are best supported by the available data. For example, we can ask whether a set of measurements $\{x_i\}$ is better described by a Gaussian or by a Cauchy distribution, or whether a set of points is better fit by a straight line or a parabola.

To find out which of two models, say M_1 and M_2, is better supported by data, we compare their posterior probabilities via the *odds ratio* in favor of model M_2 over model M_1 as

$$O_{21} \equiv \frac{p(M_2|D, I)}{p(M_1|D, I)}. \tag{5.21}$$

The posterior probability for model M (M_1 or M_2) given data D, $p(M|D, I)$ in this expression, can be obtained from the posterior pdf $p(M, \boldsymbol{\theta}|D, I)$ in eq. 5.3 using marginalization (integration) over the model parameter space spanned by $\boldsymbol{\theta}$. The posterior probability that the model M is correct given data D (a number between 0 and 1) can be derived using eqs. 5.3 and 5.4 as

$$p(M|D, I) = \frac{p(D|M, I)\, p(M|I)}{p(D|I)}, \tag{5.22}$$

where

$$E(M) \equiv p(D|M, I) = \int p(D|M, \boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}|M, I)\, d\boldsymbol{\theta} \tag{5.23}$$

is called the *marginal likelihood* for model M and it quantifies the probability that the data D would be observed *if* the model M were the correct model. In the physics literature, the marginal likelihood is often called *evidence* (despite the fact that to scientists, evidence and data mean essentially the same thing) and we adopt this term hereafter. Since the evidence $E(M)$ involves integration of the data

likelihood $p(D|M, \boldsymbol{\theta}, I)$, it is also called the *global likelihood* for model M. The global likelihood, or evidence, is a *weighted average* of the likelihood function, with the prior for model parameters acting as the weighting function.

The hardest term to compute is $p(D|I)$, but it cancels out when the odds ratio is considered:

$$O_{21} = \frac{E(M_2)\, p(M_2|I)}{E(M_1)\, p(M_1|I)} = B_{21}\, \frac{p(M_2|I)}{p(M_1|I)}. \tag{5.24}$$

The ratio of global likelihoods, $B_{21} \equiv E(M_2)/E(M_1)$, is called the *Bayes factor*, and is equal to

$$B_{21} = \frac{\int p(D|M_2, \boldsymbol{\theta}_2, I)\, p(\boldsymbol{\theta}_2|M_2, I)\, d\boldsymbol{\theta}_2}{\int p(D|M_1, \boldsymbol{\theta}_1, I)\, p(\boldsymbol{\theta}_1|M_1, I)\, d\boldsymbol{\theta}_1}. \tag{5.25}$$

The vectors of parameters, $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$, are explicitly indexed to emphasize that the two models may span vastly different parameter spaces (including the number of parameters per model).

How do we interpret the values of the odds ratio in practice? Jeffreys proposed a five-step scale for interpreting the odds ratio, where $O_{21} > 10$ represents "strong" evidence in favor of M_2 (M_2 is ten times more probable than M_1), and $O_{21} > 100$ is "decisive" evidence (M_2 is one hundred times more probable than M_1). When $O_{21} < 3$, the evidence is "not worth more than a bare mention."

As a practical example, let us consider coin flipping (this problem is revisited in detail in §5.6.2). We will compare two hypotheses; M_1: the coin has a known heads probability b_*, and M_2: the heads probability b is unknown, with a uniform prior in the range 0–1. Note that the prior for model M_1 is a delta function, $\delta(b - b_*)$. Let us assume that we flipped the coin N times, and obtained k heads. Using eq. 3.50 for the data likelihood, and assuming equal prior probabilities for the two models, it is easy to show that the odds ratio is

$$O_{21} = \int_0^1 \left(\frac{b}{b_*}\right)^k \left(\frac{1-b}{1-b_*}\right)^{N-k} db. \tag{5.26}$$

Figure 5.1 illustrates the behavior of O_{21} as a function of k for two different values of N and for two different values of b_*: $b_* = 0.5$ (M_1: the coin is fair) and $b_* = 0.1$. As this example shows, the ability to distinguish the two hypothesis improves with the sample size. For example, when $b_* = 0.5$ and $k/N = 0.1$, the odds ratio in favor of M_2 increases from $\sim$9 for $N = 10$ to $\sim$263 for $N = 20$. When $k = b_*\, N$, the odds ratio is 0.37 for $N = 10$ and 0.27 for $N = 20$. In other words, the simpler model is favored by the data, and the support strengthens with the sample size. It is easy to show by integrating eq. 5.26 that $O_{21} = \sqrt{\pi/(2N)}$ when $k = b_*\, N$ and $b_* = 0.5$. For example, to build strong evidence that a coin is fair, $O_{21} < 0.1$, it takes as many as $N > 157$ tosses. With $N = 10,000$, the heads probability of a fair coin is measured with a precision of 1% (see the discussion after eq. 3.51); the corresponding odds ratio is $O_{21} \approx 1/80$, approaching Jeffreys' decisive evidence level. Three more examples of Bayesian model comparison are discussed in §5.7.1–5.7.3.

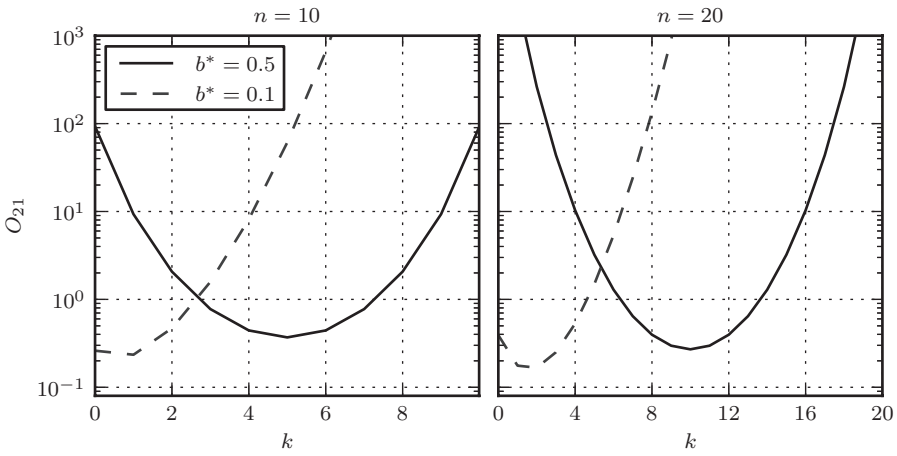

Figure 5.1. Odds ratio for two models, O_{21}, describing coin tosses (eq. 5.26). Out of N tosses (left: $N = 10$; right: $N = 20$), k tosses are heads. Model 2 is a one-parameter model with the heads probability determined from data ($b^0 = k/N$), and model 1 claims an a priori known heads probability equal to b_*. The results are shown for two values of b_*, as indicated in the legend). Note that the odds ratio is minimized and below 1 (model 1 wins) when $k = b_* N$.

5.4.1. Bayesian Hypothesis Testing

A special case of model comparison is Bayesian hypothesis testing. In this case, $M_2 = \overline{M_1}$ is a complementary hypothesis to M_1 (i.e., $p(M_1) + p(M_2) = 1$). Taking M_1 to be the "null" hypothesis, we can ask whether the data supports the alternative hypothesis M_2, i.e., whether we can reject the null hypothesis. Taking equal priors $p(M_1|I) = p(M_2|I)$, the odds ratio is

$$O_{21} = B_{21} = \frac{p(D|M_1)}{p(D|M_2)}. \tag{5.27}$$

Given that M_2 is simply a complementary hypothesis to M_1, it is not possible to compute $p(D|M_2)$ (recall that we had a well-defined alternative to M_1 in our coin example above). This inability to reject M_1 in the absence of an alternative hypothesis is very different from the hypothesis testing procedure in classical statistics (see §4.6). The latter procedure rejects the null hypothesis if it does not provide a good description of the data, that is, when it is very unlikely that the given data could have been generated as prescribed by the null hypothesis. In contrast, the Bayesian approach is based on the posterior rather than on the data likelihood, and cannot reject a hypothesis if there are no alternative explanations for observed data.

Going back to our coin example, assume we flipped the coin $N = 20$ times and obtained $k = 16$ heads. In the classical formulation, we would ask whether we can reject the null hypothesis that our coin is fair. In other words, we would ask whether $k = 16$ is a very unusual outcome (at some significance level α, say 0.05; recall §4.6) for a fair coin with $b_* = 0.5$ when $N = 20$. Using the results from §3.3.3, we find that the scatter around the expected value $k^0 = b_* N = 10$ is $\sigma_k = 2.24$. Therefore, $k = 16$ is about $2.7\sigma_k$ away from k^0, and at the adopted significance level $\alpha = 0.05$

we reject the null hypothesis (i.e., it is unlikely that $k = 16$ would have arisen by chance). Of course, $k = 16$ does not imply that it is *impossible* that the coin is fair (infrequent events happen, too!).

In the Bayesian approach, we offer an alternative hypothesis that the coin has an unknown heads probability. While this probability can be estimated from provided data (b^0), we consider *all the possible values* of b^0 when comparing the two proposed hypotheses. As shown in figure 5.1, the chosen parameters ($N = 20$, $k = 16$) correspond to the Bayesian odds ratio of ∼10 in favor of the unfair coin hypothesis.

5.4.2. Occam's Razor

The principle of selecting the simplest model that is in fair agreement with the data is known as Occam's razor. This principle was already known to Ptolemy who said, "We consider it a good principle to explain the phenomena by the simplest hypothesis possible"; see [8]. Hidden in the above expression for the odds ratio is its ability to penalize complex models with many free parameters; that is, Occam's razor is naturally included into the Bayesian model comparison.

To reveal this fact explicitly, let us consider a model $M(\boldsymbol{\theta})$, and examine just one of the model parameters, say $\mu = \theta_1$. For simplicity, let us assume that its prior pdf, $p(\mu|I)$, is flat in the range $-\Delta_\mu/2 < \mu < \Delta_\mu/2$, and thus $p(\mu|I) = 1/\Delta_\mu$. In addition, let us assume that the data likelihood can be well described by a Gaussian centered on the value of μ that maximizes the likelihood, μ^0 (see eq. 4.2), and with the width σ_μ (see eq. 4.7). When the data are much more informative than the prior, $\sigma_\mu \ll \Delta$. The integral of this approximate data likelihood is proportional to the product of σ_μ and the maximum value of the data likelihood, say $L^0(M) \equiv \max[p(D|M)]$. The global likelihood for the model M is thus approximately

$$E(M) \approx \sqrt{2\pi}\, L^0(M) \frac{\sigma_\mu}{\Delta_\mu}. \tag{5.28}$$

Therefore, $E(M) \ll L^0(M)$ when $\sigma_\mu \ll \Delta_\mu$. Each model parameter constrained by the model carries a similar multiplicative penalty, $\propto \sigma/\Delta$, when computing the Bayes factor. If a parameter, or a degenerate parameter combination, is unconstrained by the data (i.e., $\sigma_\mu \approx \Delta_\mu$), there is no penalty. The odds ratio can justify an additional model parameter only if this penalty is offset by either an increase of the maximum value of the data likelihood, $L^0(M)$, or by the ratio of prior model probabilities, $p(M_2|I)/p(M_1|I)$. If both of these quantities are similar for the two models, the one with fewer parameters typically wins.

Going back to our practical example based on coin flipping, we can illustrate how model 2 gets penalized for its free parameter. The data likelihood for model M_2 is (details are discussed in §5.6.2)

$$L(b|M_2) = C_{Nk}\, b^k\, (1-b)^{N-k}, \tag{5.29}$$

where $C_{Nk} = N!/[k!(N-k)!]$ is the binomial coefficient. The likelihood can be approximated as

$$L(b|M_2) \approx C_{Nk} \sqrt{2\pi}\, \sigma_b\, (b^0)^k\, (1-b^0)^{N-k}\, \mathcal{N}(b^0, \sigma_b) \tag{5.30}$$

with $b^0 = k/N$ and $\sigma_b = \sqrt{b^0(1-b^0)/N}$ (see §3.3.3). Its maximum is at $b = b^0$ and has the value

$$L^0(M_2) = C_{Nk}\,(b^0)^k\,(1-b^0)^{N-k}. \tag{5.31}$$

Assuming a flat prior in the range $0 \leq b \leq 1$, it follows from eq. 5.28 that the evidence for model M_2 is

$$E(M_2) \approx \sqrt{2\pi}\; L^0(M_2)\,\sigma_b. \tag{5.32}$$

Of course, we would get the same result by directly integrating $L(b|M_2)$ from eq. 5.29.

For model M_1, the approximation given by eq. 5.28 cannot be used because the prior is not flat but rather $p(b|M_1) = \delta(b - b_*)$ (the data likelihood is analogous to eq. 5.29). Instead, we can use the exact result

$$E(M_1) = C_{Nk}\,(b_*)^k\,(1-b_*)^{N-k}. \tag{5.33}$$

Hence,

$$O_{21} = \frac{E(M_2)}{E(M_1)} \approx \sqrt{2\pi}\,\sigma_b \left(\frac{b^0}{b_*}\right)^k \left(\frac{1-b^0}{1-b_*}\right)^{N-k}, \tag{5.34}$$

which is an approximation to eq. 5.26. Now we can explicitly see that the evidence in favor of model M_2 decreases (the model is "penalized") proportionally to the posterior pdf width of its free parameter. If indeed $b^0 \approx b_*$, model M_1 wins because it explained the data without any free parameter. On the other hand, the evidence in favor of M_2 increases as the data-based value b^0 becomes very different from the prior claim b_* by model M_1 (as illustrated in figure 5.1). Model M_1 becomes disfavored because it is unable to explain the observed data.

5.4.3. Information Criteria

The Bayesian information criterion (BIC, also known as the Schwarz criterion) is a concept closely related to the odds ratio, and to the Aikake information criterion (AIC; see §4.3.2 and eq. 4.17). The BIC attempts to simplify the computation of the odds ratio by making certain assumptions about the likelihood, such as Gaussianity of the posterior pdf; for details and references, see [21]. The BIC is easier to compute and, similarly to the AIC, it is based on the maximum value of the data likelihood, $L^0(M)$, rather than on its integration over the full parameter space (evidence $E(M)$ in eq. 5.23). The BIC for a given model M is computed as

$$\mathrm{BIC} \equiv -2\ln\left[L^0(M)\right] + k\,\ln N, \tag{5.35}$$

where k is the number of model parameters and N is the number of data points. The BIC corresponds to $-2\ln[E(M)]$ (to make it consistent with the AIC), and can be derived using the approximation for $E(M)$ given by eq. 5.28 and assuming $\sigma_\mu \propto 1/\sqrt{N}$.

When two models are compared, their BIC (or AIC) are compared analogously to the odds ratio, that is, the model with the smaller value wins (sometimes BIC and AIC are defined with an opposite sign, in which case the model with the largest value wins). If they are equally successful in describing the data (the first term above), then the model with fewer free parameters wins. Note that the BIC penalty for additional model parameters is stronger than for the AIC when N is sufficiently large (10–20, depending on k); very complex models are penalized more severely by the BIC than by the AIC.

Both the BIC and AIC are approximations and might not be valid if the underlying assumptions are not met (e.g., for model M_1 in our coin example above). Furthermore, unlike the odds ratio, both of them penalize unconstrained parameters. In general, it is better to compute the odds ratio when computationally feasible. We will see an example of this below in §5.8.4.

5.5. Nonuniform Priors: Eddington, Malmquist, and Lutz–Kelker Biases

In many cases, Bayesian analysis is equivalent to maximum likelihood analysis even when priors are taken into account (as we will see shortly using concrete examples discussed in §5.6). In this section, we address several important cases from astronomy where prior information *greatly affects* data interpretation. However, the same principles apply to any data set where measurement errors are not negligible and the population distribution of the measured quantity is strongly skewed, and can be easily generalized to nonastronomical contexts.

The effects of the selection function on the resulting sample are discussed in §4.9. The difference between the true distribution function $h(x)$ and its data-based estimate $f(x)$ (see eqs. 4.83 and 4.84), when caused by sample truncation, is known as the selection bias or Malmquist bias. Unfortunately, a fundamentally different effect is also often called Malmquist bias, as well as Eddington–Malmquist bias. The latter effect, addressed here, is usually encountered in brightness (magnitude) measurements and is due to the combined effects of measurement errors and nonuniform true distribution $h(x)$. Furthermore, a bias with an identical mathematical description in the context of trigonometric parallax measurements is known as Lutz–Kelker bias;[4] see [24]. The main distinguishing characteristic between the truncation (selection) bias discussed in §4.9 and the Eddington–Malmquist and Lutz–Kelker biases discussed here is that the latter two biases disappear when the measurement error for x vanishes, but the former does not.

Consider the situation illustrated in figure 5.2: an observable quantity with true values $x_{\rm true}$ is measured for a sample with true distribution $h(x_{\rm true})$. The measurements are affected by a known error distribution $e(x_{\rm obs}|x_{\rm true})$, where $x_{\rm obs}$ are the actual measured (observed) values. When $h(x_{\rm true})$ is a nonuniform distribution and measurement errors are not negligible, the distribution of $x_{\rm obs}$ for a subsample

[4] This taxonomic confusion in the literature (for an excellent summary, see [34]) apparently stems from the fact that Malmquist published the two relevant papers only two years apart (1920 and 1922) in journals that are not readily accessible today, and that those papers partially overlap with other works such as Eddington's. It may also stem from the early reluctance by astronomers to embrace Bayesian statistics (for an illuminating discussion, see [33]).

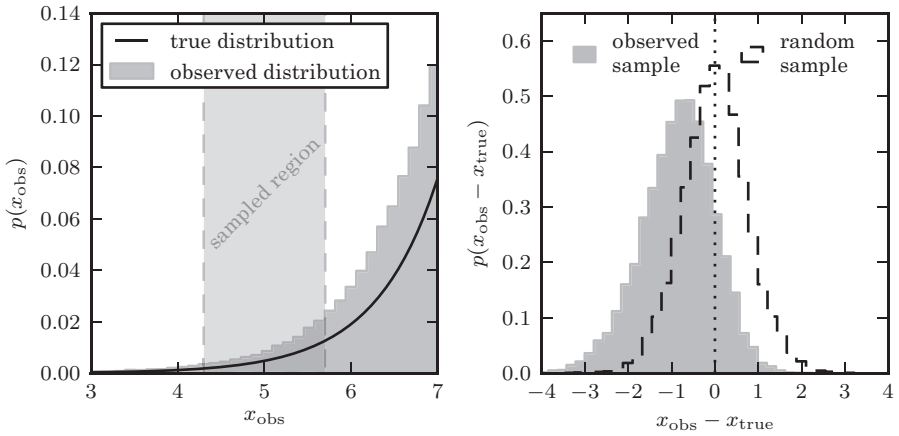

Figure 5.2. An illustration of the bias in a subsample selected using measurements with finite errors, when the population distribution is a steep function. The sample is drawn from the distribution $p(x) \propto 10^{0.6x}$, shown by the solid line in the left panel, and convolved with heteroscedastic errors with widths in the range $0.5 < \sigma < 1.5$. When a subsample is selected using "measured" values, as illustrated in the left panel, the distribution of differences between the "observed" and true values is biased, as shown by the histogram in the right panel. The distribution is biased because more objects with larger true x are scattered into the subsample from the right side, than from the left side where the true x are smaller.

selected using $x_{\rm obs}$ can be substantially different from the distribution of $x_{\rm true}$. For example, all localized features from $h(x)$, such as peaks, are "blurred" in $f(x)$ due to errors. More formally, the true and observed distributions are related via convolution (see eq. 3.44),

$$f(x) = h(x) \star e(x) = \int_{-\infty}^{\infty} h(x')\, e(x - x')\, dx'. \tag{5.36}$$

Obviously, if $e(x) = \delta(x)$ then $f(x) = h(x)$, and if $h(x) =$ constant then $f(x) =$ constant, too. This expression can be easily derived using Bayesian priors, as follows.

The prior pdf, the probability of x in the absence of measurements, is $h(x)$, the population distribution of the measured quantity. The posterior pdf, the probability of measuring value $x_{\rm obs}$, or the distribution of measured values, is

$$f(x_{\rm obs}|x_{\rm true}) \propto h(x_{\rm true})\, e(x_{\rm obs}|x_{\rm true}). \tag{5.37}$$

A given value of $x_{\rm obs}$ samples the range of $x_{\rm true}$ "allowed" by $e(x_{\rm obs}|x_{\rm true})$. When $h(x)$ is a steep rising function of x, it is more probable that $x_{\rm obs}$ corresponds to $x_{\rm true} > x_{\rm obs}$ than to $x_{\rm true} < x_{\rm obs}$, even when $e(x_{\rm obs}|x_{\rm true})$ is symmetric. Given an observed value of $x_{\rm obs}$, we do not know its corresponding $x_{\rm true}$. Nevertheless, we can still estimate $f(x_{\rm obs})$ by marginalizing over an unknown $x_{\rm true}$; that is, we integrate eq. 5.37 over $x_{\rm true}$, which yields eq. 5.36.

Figure 5.2 illustrates the bias (a systematic difference between the true and observed values) in the distribution of $x_{\rm obs}$ in samples selected by $x_{\rm obs}$ (or the full

sample, if finite). It is the act of *selecting a subsample using measured values that produces the bias*; if instead we selected a finite subsample by other means (e.g., we ask what are optical magnitudes for a subsample of radio sources detected in our optical image), the difference between measured and true magnitudes would simply follow the corresponding error distribution and not be biased, as shown in the right panel for a random subsample; for a more detailed discussion, see [32].

The example shown in figure 5.2 assumes heteroscedastic Gaussian errors. When errors are homoscedastic and Gaussian, the bias in $x_{\rm obs}$ can be computed analytically for a given $h(x)$. In this case the relationship between $f(x)$ and $h(x)$ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} h(x') \, \exp\left(\frac{-(x-x')^2}{2\sigma^2}\right) dx'. \tag{5.38}$$

Eddington was the first to show that to first order, the net effect is simply an offset between $x_{\rm obs}$ and $x_{\rm true}$, $\Delta x = x_{\rm obs} - x_{\rm true}$, and $f(x)$ is obtained by "sliding" $h(x)$ by Δx (toward smaller values of x if $h(x)$ increases with x). The offset Δx can be expressed in terms of *measured* $f(x)$ as

$$\Delta x = -\sigma^2 \, \frac{1}{f(x)} \frac{df(x)}{dx}, \tag{5.39}$$

evaluated at $x = x_{\rm obs}$. Note that Δx vanishes for $\sigma = 0$: Eddington–Malmquist and Lutz–Kelker biases become negligible for vanishing measurement errors (e.g., for a photometric error of $\sigma = 0.01$ mag, the bias becomes less than 0.0001 mag). A flat $h(x)$ (i.e., $dh(x)/dx = 0$) yields a flat $f(x)$ and thus there is no bias in this case ($\Delta x = 0$).

Special cases of much interest in astronomy correspond to very steep $h(x)$ where Δx might be nonnegligible even if σ appears small. There are two contexts, magnitude (flux) measurements and trigonometric parallax measurements, that have been frequently addressed in astronomical literature; see [33, 34]. For a spatially uniform distribution of sources, the magnitude distribution follows[5]

$$h(x) = h_o \, 10^{kx}, \tag{5.40}$$

with $k = 0.6$ (the so-called Euclidean counts), and the trigonometric parallax distribution follows[6]

$$h(x) = h_o \, x^{-p}, \tag{5.41}$$

with $p = 4$. Both of these distributions are very steep and may result in significant biases.

[5]This expression can be derived in a straightforward way using the definition of magnitude in terms of flux, and the dependence of flux on the inverse squared distance.

[6]This expression can be derived in a straightforward way using the fact that the trigonometric parallax is proportional to inverse distance.

For the case of the magnitude distribution given by eq. 5.40, the magnitude offset is

$$\Delta x = -\sigma^2 \, k \, \ln 10, \tag{5.42}$$

with the classical result for the Malmquist bias correction, $\Delta x = -1.38\sigma^2$, valid when $k = 0.6$ (measured magnitudes are brighter than true magnitudes, and inferred distances, if luminosity is known, are underestimated). As long as $\sigma \ll 1$, the offset Δx is not large compared to σ. Notably, $f(x)$ retains the same shape as $h(x)$, which justifies the use of eq. 5.40 when determining Δx using eq. 5.39. Given eq. 5.40 and the offset between $x_{\rm obs}$ and $x_{\rm true}$ given by eq. 5.42, $f(x_{\rm obs})$ is *always larger* than the corresponding $h(x_{\rm true})$: more sources with $x_{\rm obs}$ were scattered from the $x_{\rm true} > x_{\rm obs}$ range than from $x_{\rm true} < x_{\rm obs}$ (for $k > 0$).

Similarly, in the case of parallax measurements, the parallax offset is

$$\Delta x = \sigma^2 \, \frac{p}{x}. \tag{5.43}$$

The classical result for the Lutz–Kelker bias is expressed as $(x_{\rm obs}/x_{\rm true}) = 1 + p(\sigma/x)^2$; see [24]. For a constant *fractional* error (σ/x), the ratio $(x_{\rm obs}/x_{\rm true})$ is constant and thus $h(x)$ and $f(x)$ must have the same shape. For this reason, the use of $h(x)$ instead of $f(x)$ in eq. 5.39 is justified a posteriori.

If a sample is selected using parallax measurements, measured parallaxes are biased high, and implied distances are underestimated. If used for luminosity calibration, the resulting luminosity scale will be biased low. The resulting bias is as large as ∼26% even when parallax measurements have relative error as small as 20%. The fractional random luminosity error for a single measurement is twice as large as the fractional parallax error; 40% in this example. When the number of calibration sources is as large as, say, 10^4, the sample calibration error will not be 0.4% as naively expected from $\sqrt{N}$, but rather 26%! An interesting side result in the context of parallax measurements is that the parallax distribution given by eq. 5.41 is sufficiently steep that parallax measurements with relative errors exceeding about 15% ($\sigma/x > 0.15$) are practically useless in the case of individual stars: in this case the posterior pdf has a peak at 0 without another local maximum.

These classical results for bias corrections rest on assumptions that are too simplistic to be used with modern astronomical survey data sets (e.g., homoscedastic measurement errors and uniform source distribution in Euclidean geometry). For example, the magnitude measurement error is rarely constant in practice and thus the usefulness of eq. 5.42 is limited. To realistically estimate these biases, the population distribution of the measured quantity, and the dependence of measurement error on it, need to be modeled. Two practical examples illustrated in figure 5.3 simulate upcoming measurements from Gaia and LSST.

To illustrate a bias in photometric calibration, we assume that we have pairs of measurements of the same stars (e.g., from two different nights) and we compare their magnitudes. The true magnitude distribution, $h(x)$, is generated using eq. 5.40 in the range $20 < m < 25$. Gaussian heteroscedastic photometric errors are simulated using a relation expected for LSST (see [15]),

$$\sigma^2 = (0.04 - \gamma)\, x + \gamma \, x^2 \ \ (\mathrm{mag}^2), \tag{5.44}$$

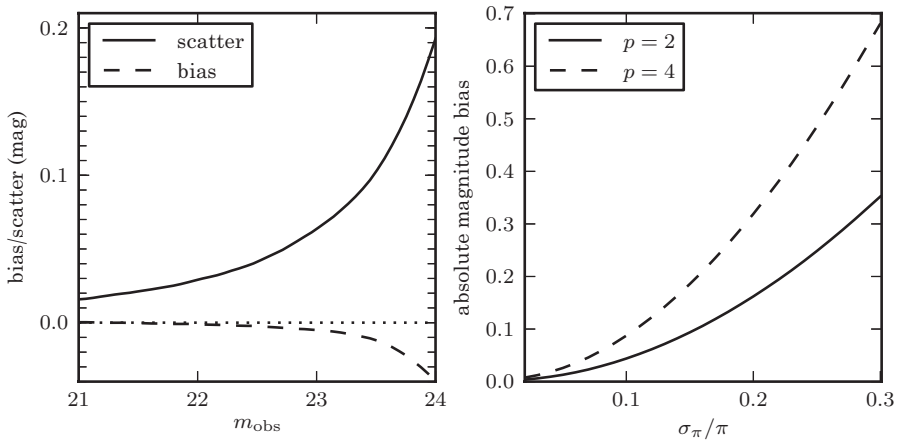

Figure 5.3. An illustration of the Eddington–Malmquist (left) and Lutz–Kelker (right) biases for mock data sets that simulate upcoming LSST and Gaia surveys (see text). The left panel shows a bias in photometric calibration when using pairs of measurements of the same stars with realistic photometric error distributions. Depending on the adopted faint limit (x-axis), the median difference between two measurements (dashed line) is biased when this limit is too close to the 5σ data limit (corresponding to errors of 0.2 mag); in this example the 5σ magnitude limit is set to 24. The solid line shows the assumed random measurement errors—if the number of stars in the sample is large, the random error for magnitude difference may become much smaller than the bias. The right panel shows the bias in absolute magnitude for samples calibrated using trigonometric parallax measurements with relative errors σ_π/π, and two hypothetical parallax distributions given by eq. 5.41 and $p = 2, 4$.

where $x = 10^{0.4\,(m-m_5)}$ and $\gamma = 0.039$ for optical sky noise-limited photometry. Sources are typically extracted from astronomical images down to the "5σ" limiting magnitude, m_5, corresponding to $\sigma = 0.2$ mag, with σ decreasing fast toward brighter magnitudes, $m < m_5$. Since σ varies with magnitude, $f(x)$ does not retain the same shape as $h(x)$. We generate two hypothetical observations, m_1 and m_2, and measure the median and the scatter for magnitude difference $\Delta m = m_1 - m_2$ as a function of magnitude limits enforced using m_1 (see the left panel in figure 5.3). Depending on the adopted faint limit, the median magnitude difference is biased when m_1 is too close to m_5.

In order to minimize the impact of the Eddington–Malmquist bias, the faint limit of the sample should be at least a magnitude brighter than m_5. Alternatively, if one of the two sets of measurements has a much fainter m_5 than the other one, then the sample should be selected using that set.

To illustrate a bias in calibration of absolute magnitudes using parallax measurements, we simulate an LSST–Gaia sample. The magnitude distribution is generated using eq. 5.40 in the r magnitude range $17 < r < 20$ that roughly corresponds to the brightness overlap between Gaia (faint limit at $r \sim 20$) and LSST (saturates at $r \sim 17$). Trigonometric parallax errors are simulated using a relation similar to that expected for Gaia,

$$\sigma_\pi = 0.3 \times 10^{0.2\,(r-20)} \ \text{milliarcsec}, \tag{5.45}$$

and we consider a hypothetical sample of stars whose absolute magnitudes are about 10. We compute a bias in absolute magnitude measurement as

$$\Delta M = 5 \log_{10} \left[1 + 4 \left(\frac{\sigma_\pi}{\pi} \right)^2 \right], \tag{5.46}$$

where π is the parallax corresponding to a star with given r and absolute magnitude $M_r = 10$, and we analyze two different parallax distributions described by $p = 2$ and $p = 4$ (see eq. 5.41). As illustrated in the right panel in figure 5.3, in order to minimize this bias below, say, 0.05 mag, the sample should be selected by $\sigma_\pi / \pi < 0.1$.

5.6. Simple Examples of Bayesian Analysis: Parameter Estimation

In this section we illustrate the important aspects of Bayesian parameter estimation using specific practical examples. In the following section (§5.7), we will discuss several examples of Bayesian model selection. The main steps for Bayesian parameter estimation and model selection are summarized in §5.1.3.

5.6.1. Parameter Estimation for a Gaussian Distribution

First, we will solve a simple problem where we have a set of N measurements, $\{x_i\}$, of, say, the length of a rod. The measurement errors are Gaussian, and the measurement error for each measurement is known and given as σ_i (heteroscedastic errors). We seek the posterior pdf for the length of the rod, μ: $p(\mu|\{x_i\}, \{\sigma_i\})$.

Given that the likelihood function for a single measurement, x_i, is *assumed* to follow a Gaussian distribution (see below for a generalization), the likelihood for obtaining data $D = \{x_i\}$ given μ (and $\{\sigma_i\}$) is simply the product of likelihoods for individual data points,

$$p(\{x_i\}|\mu, I) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma_i^2} \right). \tag{5.47}$$

For the prior for μ, we shall adopt the least informative choice: a uniform distribution over some very wide interval ranging from $\mu_{\rm min}$ to $\mu_{\rm max}$:

$$p(\mu|I) = C, \quad \text{for } \mu_{\rm min} < \mu < \mu_{\rm max}, \tag{5.48}$$

where $C = (\mu_{\rm max} - \mu_{\rm min})^{-1}$, and 0 otherwise (the choice of $\mu_{\rm min}$ and $\mu_{\rm max}$ will not have a major impact on the solution: for example, in this case we could assume $\mu_{\rm min} = 0$ and $\mu_{\rm max}$ is the radius of Earth). The logarithm of the posterior pdf for μ is then

$$L_p = \ln\left[p(\mu|\{x_i\}, \{\sigma_i\}, I) \right] = \text{constant} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma_i^2} \tag{5.49}$$

(remember that we use L and lnL as notation for the data likelihood and its logarithm; L_p is reserved for the logarithm of the posterior pdf). This result is

analogous to eq. 4.8 and the only difference is in the value of the constant term (due to the prior for μ and ignoring the $p(D|I)$ term), which is unimportant in this analysis.

Again, as a result of the Gaussian error distribution we can derive an analytic solution for the maximum likelihood estimator of μ by setting $\left(dL_p/d\mu\right)|_{(\mu=\mu_0)} = 0$,

$$\mu_0 = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}, \tag{5.50}$$

with weights $w_i = \sigma_i^{-2}$. That is, μ_0 is simply a weighted arithmetic mean of all measurements. When all σ_i are equal, we obtain the standard result given by eq. 3.31 and eq. 4.5. The posterior pdf for μ is a Gaussian centered on μ_0, and with the width given by (in analogy with eq. 4.6)

$$\sigma_\mu = \left(-\frac{d^2 L_p}{d\mu^2}\,|_{\mu=\mu_0}\right)^{-1/2} = \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right)^{-1/2} = \left(\sum_i w_i\right)^{-1/2}. \tag{5.51}$$

Because we used a flat prior for $p(M|I)$, these results are identical to those that follow from the maximum likelihood method. Note that although eq. 5.51 is only an approximation based on quadratic Taylor expansion of the logarithm of the posterior pdf, it is exact in this case because there are no terms higher than μ^2 in eq. 5.49. Again, when all σ_i are equal to σ, we obtain the standard result given by eqs. 3.34 and 4.7. The key conclusion is that the posterior pdf for μ is *Gaussian in cases when σ_i are known*, regardless of the data set size N. This is not true when σ is unknown and also determined from data, as follows.

Let us now solve a similar but more complicated problem when a set of N values (measurements), $\{x_i\}$, is drawn from an unspecified Gaussian distribution, $\mathcal{N}(\mu, \sigma)$. That is, here *σ also needs to be determined from data;* for example, it could be that individual measurement errors are always negligible compared to the intrinsic spread σ of the measured quantity (e.g., when measuring the weight of a sample of students with a microgram precision), or that all measurements of a rod do have the same unknown precision.

We seek the two-dimensional posterior pdf $p(\mu, \sigma|\{x_i\})$. This problem is frequently encountered in practice and the most common solution for estimators of μ and σ is given by eqs. 3.31 and 3.32. Much less common is the realization that the assumption of the Gaussian uncertainty of μ, with its width given by eq. 3.34, is valid *only in the large N limit* when σ is not known a priori. When N is not large, the posterior pdf for μ follows Student's t distribution. Here is a how to derive this result using the Bayesian framework.

Given that the likelihood function for a single measurement, x_i, is assumed to follow a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, the likelihood for all measurements is given by

$$p(\{x_i\}|\mu, \sigma, I) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right). \tag{5.52}$$

This equation is identical to eq. 4.2. However, the main and fundamental difference is that σ in eq. 4.2 was assumed to be known, while σ in eq. 5.52 is to be estimated, thus making the posterior pdf a function of two parameters (similarly, σ_i in eq. 5.47 were also assumed to be known).

We shall again adopt a uniform prior distribution for the location parameter μ and, following the discussion in §5.2, a uniform prior distribution for $\ln\sigma$, which leads to

$$p(\mu, \sigma|I) \propto \frac{1}{\sigma}, \quad \text{for } \mu_{\rm min} \le \mu \le \mu_{\rm max} \text{ and } \sigma_{\rm min} \le \sigma \le \sigma_{\rm max}. \tag{5.53}$$

The exact values of the distribution limits are not important as long as they do not significantly truncate the likelihood. However, because we will need a properly normalized pdf in the context of the model comparison discussed in §5.7.1, we explicitly write the full normalization here:

$$p(\{x_i\}|\mu, \sigma, I)\, p(\mu, \sigma|I) = C\, \frac{1}{\sigma^{(N+1)}} \prod_{i=1}^{N} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right), \tag{5.54}$$

where

$$C = (2\pi)^{-N/2}\, (\mu_{\rm max} - \mu_{\rm min})^{-1} \left[\ln\left(\frac{\sigma_{\rm max}}{\sigma_{\rm min}}\right)\right]^{-1}. \tag{5.55}$$

The value of C can be a very small number, especially when N is large. For example, with ($\mu_{\rm min} = -10$, $\mu_{\rm max} = 10$, $\sigma_{\rm min} = 0.01$, $\sigma_{\rm max} = 100$, $N = 10$), then $C = 5.6 \times 10^{-7}$.

The logarithm of the posterior pdf now becomes (cf. eq. 5.49)

$$L_p \equiv \ln\left[p(\mu, \sigma|\{x_i\}, I)\right] = \text{constant} - (N+1)\, \ln\sigma - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}. \tag{5.56}$$

Had we assumed a uniform distribution of σ, instead of $\ln\sigma$, then the factor multiplying $\ln\sigma$ would change from $(N+1)$ to N. Using the equality

$$\sum_{i=1}^{N}(x_i-\mu)^2 = N\,(\overline{x}-\mu)^2 + \sum_{i=1}^{N}(x_i-\overline{x})^2 \tag{5.57}$$

and with the substitution $V = N^{-1}\sum_{i=1}^{N}(x_i-\overline{x})^2$ (note that $V = (N-1)s^2/N$, where s is the sample standard deviation given by eq. 3.32), we can rewrite eq. 5.56 in terms of three data-based quantities, N, $\overline{x}$ and V:

$$L_p = \text{constant} - (N+1)\, \ln\sigma - \frac{N}{2\sigma^2}\left((\overline{x}-\mu)^2 + V\right). \tag{5.58}$$

Note that irrespective of the size of our data set we only need these three numbers (N, $\overline{x}$, and V) to *fully capture its entire information content* (because we assumed

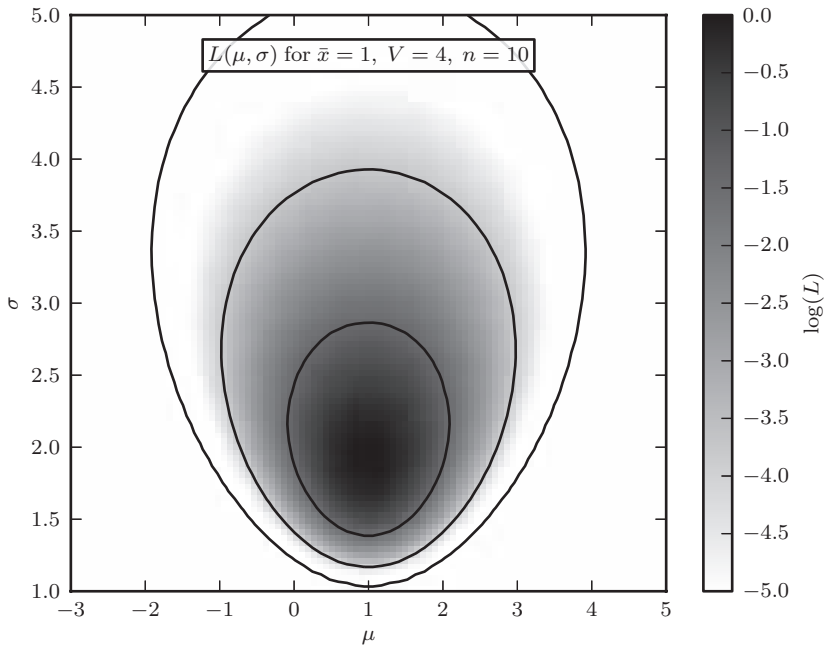

Figure 5.4. An illustration of the logarithm of the posterior probability density function for μ and σ, $L_p(\mu, \sigma)$ (see eq. 5.58) for data drawn from a Gaussian distribution and $N = 10, \overline{x} = 1$, and $V = 4$. The maximum of L_p is renormalized to 0, and color coded as shown in the legend. The maximum value of L_p is at $\mu_0 = 1.0$ and $\sigma_0 = 1.8$. The contours enclose the regions that contain 0.683, 0.955, and 0.997 of the cumulative (integrated) posterior probability.

a Gaussian likelihood function). They are called *sufficient statistics* because they summarize all the information in the data that is relevant to the problem (for formal definitions see Wass10).

An illustration of $L_p(\mu, \sigma)$ for $N = 10, \overline{x} = 1$, and $V = 4$ is shown in figure 5.4. The position of its maximum (μ_o, σ_o) can be found using eqs. 4.3 and 5.58: $\mu_0 = \overline{x}$ and $\sigma_0^2 = VN/(N+1)$ (i.e., σ_0 is equal to the sample standard deviation; see eq. 3.32).

The region of the (μ, σ) plane which encloses a given cumulative probability for the posterior pdf (or regions in the case of multimodal posterior pdfs) can be found by the following simple numerical procedure. The posterior pdf up to a normalization constant is simply $\exp(L_p)$ (e.g., see eqs. 5.56 or 5.58). The product of $\exp(L_p)$ and the pixel area (assuming sufficiently small pixels so that no interpolation is necessary) can be summed up to determine the normalization constant (the integral of the posterior pdf over all model parameters must be unity). The renormalized posterior pdf can be sorted, while keeping track of the corresponding pixel for each value, and the corresponding cumulative distribution computed. Given a threshold of p%, all the pixels for which the cumulative probability is larger than $(1 - p/100)$ will outline the required region.

For a given $\sigma = \sigma'$, the maximum value of L_p is found along the $\mu = \mu_0 = \overline{x}$ line, and the posterior probability $p(\mu|\sigma = \sigma')$ is a Gaussian (same result as given by eqs. 4.5 and 4.7). However, now we do not know the true value of σ. When deriving the posterior probability $p(\mu)$, we need to marginalize (integrate) over all

possible values of σ (in figure 5.4, at each value of μ, we "sum all the pixels" in the corresponding vertical slice through the image and renormalize the result),

$$p(\mu|\{x_i\}, I) = \int_0^\infty p(\mu, \sigma|\{x_i\}, I)\, d\sigma, \tag{5.59}$$

yielding (starting with eq. 5.58 and using the substitution $t = 1/\sigma$ and integration by parts)

$$p(\mu|\{x_i\}, I) \propto \left[1 + \frac{(\overline{x} - \mu)^2}{V}\right]^{-N/2}. \tag{5.60}$$

It is easy to show that this result corresponds to Student's t distribution (see eq. 3.60) with $k = N - 1$ degrees of freedom for the variable $t = (\overline{x} - \mu)/(s/\sqrt{N})$, where the sample standard deviation s is given by eq. 3.32. Had we assumed a uniform prior for σ, we would have obtained Student's t distribution with $k = (N - 2)$ degrees of freedom (the difference between these two solutions becomes negligible for large N).

The posterior marginal pdf $p(\mu|\{x_i\}, I)$ for $N = 10$, $\overline{x} = 1$, and $V = 4$ is shown in figure 5.5, for both σ priors (note that the uniform prior for σ gives a slightly wider posterior pdf). While the core of the distribution is similar to a Gaussian with parameters given by eqs. 3.31 and 3.34, its tails are much more extended. As N increases, Student's t distribution becomes more similar to a Gaussian distribution and thus $p(\mu|\{x_i\}, I)$ eventually becomes Gaussian, as expected from the central limit theorem.

Analogously to determining $p(\mu|\{x_i\}, I)$, the posterior pdf for σ is derived using marginalization,

$$p(\sigma|\{x_i\}, I) = \int_0^\infty p(\mu, \sigma|\{x_i\}, I)\, d\mu, \tag{5.61}$$

yielding (starting with eq. 5.58 again)

$$p(\sigma|\{x_i\}, I) \propto \frac{1}{\sigma^N} \exp\left(\frac{-NV}{2\sigma^2}\right). \tag{5.62}$$

Had we assumed a uniform prior for σ, the first term would have been $1/\sigma^{(N-1)}$. Eq. 5.62 is equivalent to the χ^2 distribution with $k = N - 1$ degrees of freedom for variable $Q = NV/\sigma^2$ (see eq. 3.58). An analogous result is known as Cochran's theorem in classical statistics. For a uniform prior for σ, the number of degrees of freedom is $k = N + 1$.

The posterior marginal pdf $p(\sigma|\{x_i\}, I)$ for our example is shown in figure 5.5. As is easily discernible, the posterior pdf for σ given by eq. 5.62 is skewed and not Gaussian, although the standard result given by eq. 3.35 implies the latter. The result from eq. 3.35 can be easily derived from eq. 5.62 using the approximation given by eq. 4.6, and is also shown in figure 5.5 (eq. 3.35 corresponds to a uniform σ prior; for a prior proportional to σ^{-1}, there is an additional $(N - 1)/(N + 1)$ multiplicative term).

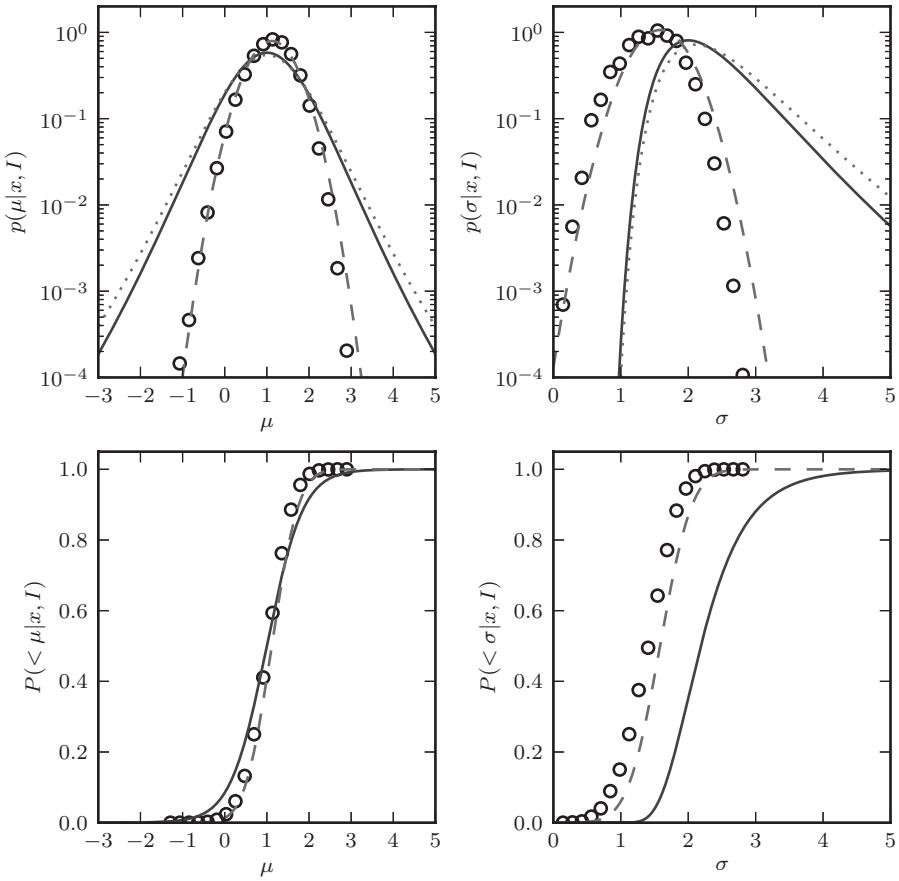

Figure 5.5. The solid line in the top-left panel shows the posterior probability density function $p(\mu|\{x_i\}, I)$ described by eq. 5.60, for $N = 10$, $\overline{x} = 1$ and $V = 4$ (integral over σ for the two-dimensional distribution shown in figure 5.4). The dotted line shows an equivalent result when the prior for σ is uniform instead of proportional to σ^{-1}. The dashed line shows the Gaussian distribution with parameters given by eqs. 3.31 and 3.34. For comparison, the circles illustrate the distribution of the bootstrap estimates for the mean given by eq. 3.31. The solid line in the top-right panel shows the posterior probability density function $p(\sigma|\{x_i\}, I)$ described by eq. 5.62 (integral over μ for the two-dimensional distribution shown in figure 5.4). The dotted line shows an equivalent result when the prior for σ is uniform. The dashed line shows a Gaussian distribution with parameters given by eqs. 3.32 and 3.35. The circles illustrate the distribution of the bootstrap estimates for σ given by eq. 3.32. The bottom two panels show the corresponding cumulative distributions for solid and dashed lines, and for bootstrap estimates, from the top panel.

When N is small, the posterior probability for large σ is much larger than that given by the Gaussian approximation. For example, the cumulative distributions shown in the bottom-right panel in figure 5.5 indicate that the probability of $\sigma > 3$ is in excess of 0.1, while the Gaussian approximation gives $\sim$0.01 (with the discrepancy increasing fast for larger σ). Sometimes, this inaccuracy of the Gaussian approximation can have a direct impact on derived scientific conclusions. For example, assume that we measured velocity dispersion (i.e., the sample standard deviation)

of a subsample of 10 stars from the Milky Way's halo and obtained a value of 50 km s^{-1}. The Gaussian approximation tells us that (within the classical framework) we can reject the hypothesis that this subsample came from a population with velocity dispersion greater than 85 km s^{-1} (representative of the halo population) at a highly significant level (p value $\sim$ 0.001) and thus we might be tempted to argue that we discovered a stellar stream (i.e., a population with much smaller velocity dispersion). However, eq. 5.62 tells us that a value of 85 km s^{-1} or larger cannot be rejected even at a generous $\alpha = 0.05$ significance level! In addition, the Gaussian approximation and classical framework formally allow $\sigma \leq 0$, an impossible conclusion which is easily avoided by adopting a proper prior in the Bayesian approach. That is, the problem with negative s that we mentioned in the context of eq. 3.35 is resolved when using eq. 5.62. Therefore, when N is small (less than 10, though $N < 100$ is a safe bet for most applications), the confidence interval (i.e., credible region in the Bayesian framework) for σ should be evaluated using eq. 5.62 instead of eq. 3.35.

For a comparison of classical and Bayesian results, figure 5.5 also shows bootstrap confidence estimates for μ and σ (circles). As is evident, when the sample size is small, they have unreliable (narrower) tails and are more similar to Gaussian approximations with the widths given by eqs. 3.34 and 3.35. Similar widths are obtained using the jackknife method, but in this case we would use Student's t distribution with $N - 1$ degrees of freedom (see §4.5). The agreement with the above posterior marginal probability distributions would be good in the case of μ, but the asymmetric behavior for σ would not be reproduced. Therefore, as discussed in §4.5, the bootstrap and jackknife methods should be used with care.

Gaussian distribution with Gaussian errors

The posterior pdfs for μ and σ given by eqs. 5.60 and 5.62 correspond to a case where $\{x_i\}$ are drawn from an unspecified Gaussian distribution, $\mathcal{N}(\mu, \sigma)$. The width σ can be interpreted in two ways: it could correspond to the *intrinsic* spread σ of the measured quantity when measurement errors are always negligible, or it could simply be the unknown homoscedastic measurement error when measuring a single-valued quantity (such as the length of a rod in the above examples). A more general case is when the measured quantity is drawn from some distribution whose parameters we are trying to estimate, and the known measurement errors are heteroscedastic. For example, we might be measuring the radial velocity dispersion of a stellar cluster using noisy estimates of radial velocity for individual stars.

If the errors are homoscedastic, the resulting distribution of measurements is Gaussian: this is easily shown by recognizing the fact that the sum of two random variables has a distribution equal to the convolution of the input distributions, and that the convolution of two Gaussians is itself a Gaussian. However, when errors are heteroscedastic, the resulting distribution of measurements is not itself a Gaussian. As an example, figure 5.6 shows the pdf for the $\mathcal{N}(0, 1)$ distribution sampled with heteroscedastic Gaussian errors with widths uniformly distributed between 0 and 3. The Anderson–Darling statistic (see §4.7.4) for the resulting distribution is $A^2 = 3088$ (it is so large because N is large), strongly indicating that the data are not drawn from a normal distribution. The best-fit normal curves (based on both the sample variance and interquartile range) are shown for comparison.

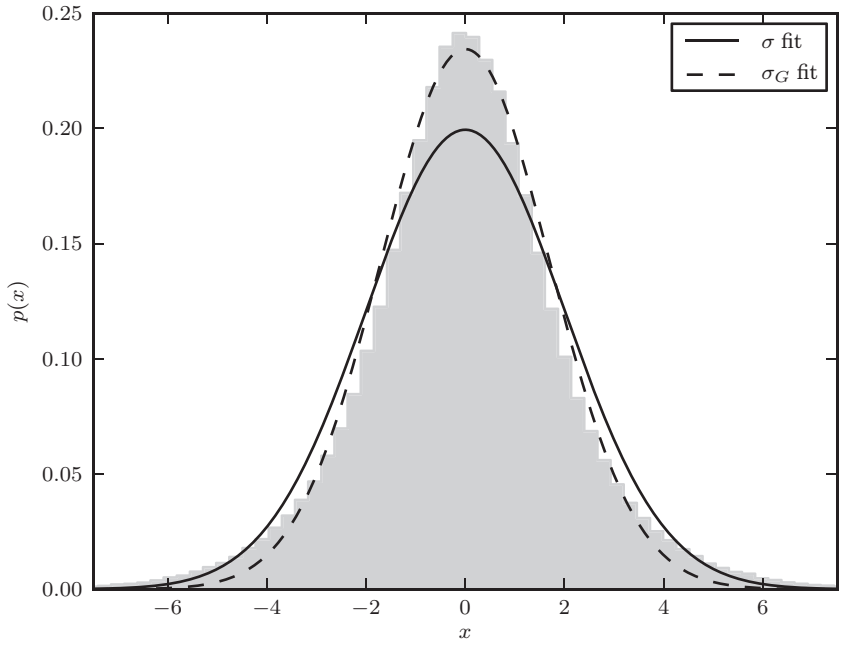

Figure 5.6. The distribution of 10^6 points drawn from $\mathcal{N}(0, 1)$ and sampled with heteroscedastic Gaussian errors with widths, e_i, uniformly distributed between 0 and 3. A linear superposition of these Gaussian distributions with widths equal to $\sqrt{1 + e_i^2}$ results in a non-Gaussian distribution. The best-fit Gaussians centered on the sample median with widths equal to sample standard deviation and quartile-based σ_G (eq. 3.36) are shown for comparison.

In order to proceed with parameter estimation in this situation, we shall assume that data were drawn from an intrinsic $\mathcal{N}(\mu, \sigma)$ distribution, and that measurement errors are also Gaussian and described by the known width e_i. Starting with an analog of eq. 5.52,

$$p(\{x_i\}|\mu, \sigma, I) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}(\sigma^2 + e_i^2)^{1/2}} \exp\left(\frac{-(x_i - \mu)^2}{2(\sigma^2 + e_i^2)}\right), \tag{5.63}$$

and following the same steps as above (with uniform priors for μ and σ), we get an analog of eq. 5.56:

$$L_p = \text{constant} - \frac{1}{2}\sum_{i=1}^{N}\left(\ln(\sigma^2 + e_i^2) + \frac{(x_i - \mu)^2}{\sigma^2 + e_i^2}\right). \tag{5.64}$$

However, in this case σ is "coupled" to e_i and must remain inside the sum, thus preventing us from capturing the entire information content of our data set using only 3 numbers (N, $\overline{x}$ and V), as we did when $e_i \ll \sigma$ (see eq. 5.58). This difficulty arises because the underlying distribution of $\{x_i\}$ is no longer Gaussian—instead it is a weighted sum of Gaussians with varying widths (recall figure 5.6; of course, the likelihood function for a single measurement is Gaussian). Compared to a Gaussian

with the same σ_G (see eq. 3.36), the distribution of $\{x_i\}$ has more pronounced tails that reflect the distribution of e_i with finite width.

By setting the derivative of L_p with respect to μ to zero, we can derive an analog of eq. 5.50,

$$\mu_0 = \frac{\sum_i w_i x_i}{\sum_i w_i}, \tag{5.65}$$

except that weights are now

$$w_i = \frac{1}{\sigma_0^2 + e_i^2}. \tag{5.66}$$

These weights are fundamentally different from the case where σ (or σ_i) is known: σ_0 is now a quantity we are trying to estimate! By setting the derivative of L_p with respect to σ to zero, we get the constraint

$$\sum_{i=1}^{N} \frac{1}{\sigma_0^2 + e_i^2} = \sum_{i=1}^{N} \frac{(x_i - \mu_0)^2}{(\sigma_0^2 + e_i^2)^2}. \tag{5.67}$$

Therefore, we cannot obtain a closed-form expression for the MAP estimate of σ_0. In order to obtain MAP estimates for μ and σ, we need to solve the system of two complicated equations, eqs. 5.65 and 5.67. We have encountered a very similar problem when discussing the expectation maximization algorithm in §4.4.3. A straightforward way to obtain solutions is an iterative procedure: we start with a guess for μ_0 and σ_0 and obtain new estimates from eq. 5.65 (trivial) and eq. 5.67 (needs to be solved numerically).

Of course, there is nothing to stop us from simply evaluating L_p given by eq. 5.64 on a grid of μ and σ, as we did earlier in the example illustrated in figure 5.4. We generate a data set using $N = 10$, $\mu = 1$, $\sigma = 1$, and errors $0 < e_i < 3$ drawn from a uniform distribution. This e_i distribution is chosen to produce a similar sample variance as in the example from figure 5.4 ($V \approx 4$). Once e_i is generated, we draw a data value from a Gaussian distribution centered on μ and width equal to $(\sigma^2+e_i^2)^{1/2}$. The resulting posterior pdf is shown in figure 5.7. Unlike the case with homoscedastic errors (see figure 5.4), the posterior pdf in this case is not symmetric with respect to the $\mu = 1$ line.

In practice, approximate estimates of μ_0 and σ_0 can be obtained without the explicit computation of the posterior pdf. Numerical simulations show that the sample median is an efficient and unbiased estimator of μ_0 (by symmetry), and its uncertainty can be estimated using eq. 3.34, with the standard deviation replaced by the quartile-based width estimator given by eq. 3.36, σ_G. With a data-based estimate of σ_G and the median error e_{50}, σ_0 can be estimated as

$$\sigma_0^2 = \zeta^2 \sigma_G^2 - e_{50}^2 \tag{5.68}$$

(to avoid solutions consistent with zero when N is small, this should be used only for large N, say $N > 100$). Here, ζ depends on the details of the distribution of e_i and

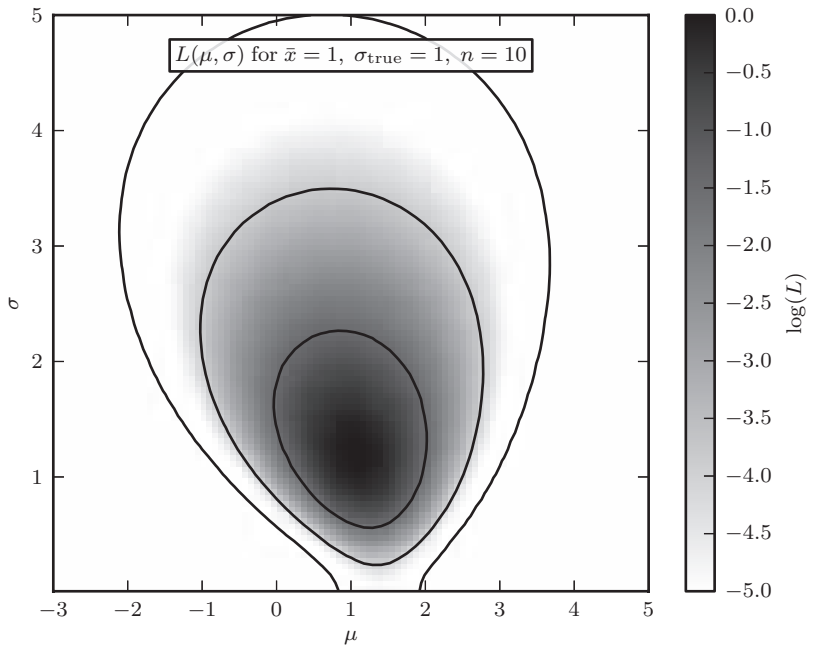

Figure 5.7. The logarithm of the posterior probability density function for μ and σ, $L_p(\mu, \sigma)$, for a Gaussian distribution with heteroscedastic Gaussian measurement errors (sampled uniformly from the 0–3 interval), given by eq. 5.64. The input values are $\mu = 1$ and $\sigma = 1$, and a randomly generated sample has 10 points. Note that the posterior pdf is not symmetric with respect to the $\mu = 1$ line, and that the outermost contour, which encloses the region that contains 0.997 of the cumulative (integrated) posterior probability, allows solutions with $\sigma = 0$.

can be estimated as

$$\zeta = \frac{\text{median}(\tilde{\sigma_i})}{\text{mean}(\tilde{\sigma_i})}, \tag{5.69}$$

where

$$\tilde{\sigma_i} = (\sigma_G^2 + e_i^2 - e_{50}^2)^{1/2}. \tag{5.70}$$

If all $e_i = e$, then $\zeta = 1$ and $\sigma_0^2 = \sigma_G^2 - e^2$, as expected from the convolution of two Gaussians. Of course, if $e_i \ll \sigma_G$ then $\zeta \to 1$ and $\sigma_0 \to \sigma_G$ (i.e., $\{x_i\}$ are drawn from $\mathcal{N}(\mu, \sigma_0)$).

These closed-form solutions follow from the result that σ_G for a weighted sum of Gaussians $\mathcal{N}(\mu, \sigma)$ with varying σ is *approximately* equal to the mean value of σ (i.e., mean$[(\sigma_o^2 + e_i^2)^{1/2}] = \sigma_G$), and the fact that the median of a random variable which is a function of another random variable is equal to the value of that function evaluated for the median of the latter variable (i.e., median$(\sigma_o^2 + e_i^2) = \sigma_o^2 + e_{50}^2$). For very large samples ($N > 1000$), the error for σ_0 given by eq. 5.68 becomes smaller than its bias (10–20%; that is, this estimator is not consistent). If this bias level is important in a specific application, a quick remedy is to compute the posterior pdf

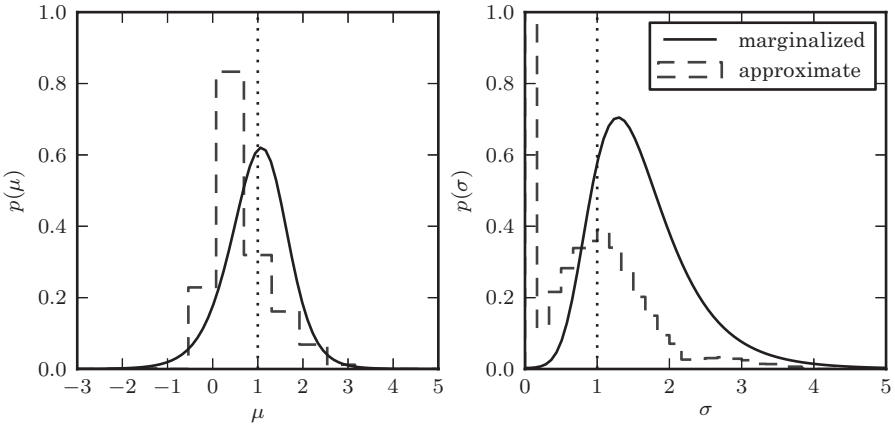

Figure 5.8. The solid lines show marginalized posterior pdfs for μ (left) and σ (right) for a Gaussian distribution with heteroscedastic Gaussian measurement errors (i.e., integrals over σ and μ for the two-dimensional distribution shown in figure 5.7). For comparison, the dashed histograms show the distributions of approximate estimates for μ and σ (the median and given by eq. 5.68, respectively) for 10,000 bootstrap resamples of the same data set. The true values of μ and σ are indicated by the vertical dotted lines.

given by eq. 5.64 in the neighborhood of approximate solutions and find the true maximum.

These "poor man's" estimators for the example discussed here ($\zeta = 0.94$) are also shown in figure 5.8. Because the sample size is fairly small ($N = 10$), in a few percent of cases $\sigma_0 < 0.1$. Although these estimators are only approximate, they are a *much better* solution than to completely ignore σ_0 and use, for example, the weighted mean formula (eq. 5.50) with $w_i = 1/e_i^2$. The main reason for this better performance is that here $\{x_i\}$ does not follow a Gaussian distribution (although both the intrinsic distribution of the measured quantity is Gaussian *and* measurement errors are Gaussian). Of course, the best option is to use the full Bayesian solution.

We have by now collected a number of different examples based on Gaussian distributions. They are summarized in table 5.1. The last row addresses the seemingly hopeless problem when both σ_0 and the heteroscedastic errors $\{e_i\}$ are unknown. Nevertheless, even in this case data contain information and can be used to place an upper limit $\sigma_0 \leq \sigma_G$ (in addition to using the median to estimate μ). Furthermore, $\{e_i\}$ can be considered model parameters, too, and marginalized over with some reasonable priors (e.g., a flat prior between 0 and the maximum value of $|x_i - \mu|$) to derive a better constraint for σ_0. This is hard to do analytically, but easy numerically, and we shall address this case in §5.8.5.

5.6.2. Parameter Estimation for the Binomial Distribution

We have already briefly discussed the coin flip example in §5.4. We revisit it here in the more general context of parameter estimation for the binomial distribution. Given a set of N measurements (or trials), $\{x_i\}$, drawn from a binomial distribution described with parameter b (see §3.3.3), we seek the posterior probability distribution $p(b|\{x_i\})$. Similarly to the Gaussian case discussed above, when N is large, b and its

TABLE 5.1.
Summary of the results for estimating parameters of $\mathcal{N}(\mu, \sigma_0)$ when data $\{x_i\}$ have a Gaussian error distribution given by $\{e_i\}$ (each data point is drawn from $\mathcal{N}(\mu, \sigma_i)$ with $\sigma_i^2 = \sigma_0^2 + e_i^2$). Weights w_i refer to eqs. 5.50 and 5.51, and s is the standard deviation given by eq. 3.32. If the error distribution is homoscedastic, but not necessarily Gaussian, use the median instead of the weighted mean, and the quartile-based width estimate σ_G (eq. 3.36) instead of the standard deviation.

σ_0	$\{e_i\}$	Weights	Description
σ_0	$e_i = e$	$w_i = 1$	homoscedastic, both σ_0 and e are known
σ_0	$e_i = 0$	$w_i = 1$	homoscedastic, errors negligible, $\sigma_0 = s$
0	$e_i = e$	$w_i = 1$	homoscedastic, single-valued quantity, $e = s$
σ_0	$e_i = e$	$w_i = 1$	homoscedastic, e known, $\sigma_0^2 = (s^2 - e^2)$
0	e_i known	$w_i = e_i^{-2}$	errors heteroscedastic but assumed known
σ_0	e_i known	no closed form	σ_0 unknown; see eq. 5.64 and related discussion
σ_0	unknown	no closed form	upper limit for σ_0; numerical modeling; see text (also §5.8.5).

(presumably Gaussian) uncertainty can be determined as discussed in §3.3.3. For small N, the proper procedure is as follows.

Here the data set $\{x_i\}$ is discrete: all outcomes are either 0 (heads) or 1 (tails, which we will consider "success"). An astronomical analog might be the computation of the fraction of galaxies which show evidence for a black hole in their center. Given a model parametrized by the probability of success b, the likelihood that the data set contains k outcomes equal to 1 is given by eq. 3.50. Assuming that the prior for b is flat in the range 0–1, the posterior probability for b is

$$p(b|k, N) = C\, b^k\, (1 - b)^{N-k}, \tag{5.71}$$

where k is now the actual observed number of successes in a data set of N values, and the normalization constant C can be determined from the condition $\int_0^1 p(b|k, N)\, db = 1$ (alternatively, we can make use of the fact that the beta distribution is a conjugate prior for binomial likelihood; see §5.2.3). The maximum posterior occurs at $b_0 = k/N$.

For a concrete numerical example, let us assume that we studied $N = 10$ galaxies and found a black hole in $k = 4$ of them. Our best estimate for the fraction of galaxies with black holes is $b_0 = k/N = 0.4$. An interesting question is, "What is the probability that, say, $b < 0.1$?" For example, your colleague's theory placed an upper limit of 10% for the fraction of galaxies with black holes and you want to test this theory using classical framework ("Can it be rejected at a confidence level $\alpha = 0.01$?").

Using the Gaussian approximation discussed in §3.3.3, we can compute the standard error for b_0 as

$$\sigma_b = \left[\frac{b_0\,(1 - b_0)}{N}\right]^{1/2} = 0.155, \tag{5.72}$$

and conclude that the probability for $b < 0.1$ is ∼0.03 (the same result follows from eq. 4.6). Therefore, at a confidence level $\alpha = 0.01$ the theory is not rejected. However,

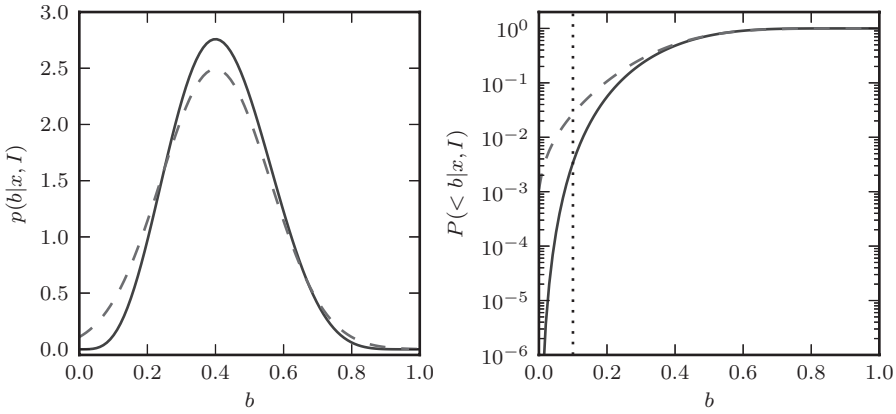

Figure 5.9. The solid line in the left panel shows the posterior pdf $p(b|k, N)$ described by eq. 5.71, for $k = 4$ and $N = 10$. The dashed line shows a Gaussian approximation described in §3.3.3. The right panel shows the corresponding cumulative distributions. A value of 0.1 is marginally likely according to the Gaussian approximation ($p_{\rm approx}(< 0.1) \approx 0.03$) but strongly rejected by the true distribution ($p_{\rm true}(< 0.1) \approx 0.003$).

the exact solution given by eq. 5.71 and shown in figure 5.9 is not a Gaussian! By integrating eq. 5.71, you can show that $p(b < 0.1|k = 4, N = 10) = 0.003$, and therefore your data do reject your colleague's theory[7] (note that in the Bayesian framework we need to specify an alternative hypothesis; see §5.7). When N is not large, or b_0 is close to 0 or 1, one should avoid using the Gaussian approximation when estimating the credible region (or the confidence interval) for b.

5.6.3. Parameter Estimation for the Cauchy (Lorentzian) Distribution

As already discussed in §3.3.5, the mean of a sample drawn from the Cauchy distribution is not a good estimator of the distribution's location parameter. In particular, the mean value for many independent samples will themselves follow the same Cauchy distribution, and will not benefit from the central limit theorem (because the variance does not exist). Instead, the location and scale parameters for a Cauchy distribution (μ and γ) can be simply estimated using the median value and interquartile range for $\{x_i\}$. We shall now see how we can estimate the parameters of a Cauchy distribution using a Bayesian approach.

As a practical example, we will use the lighthouse problem due to Gull and discussed in Siv06 (a mathematically identical problem is also discussed in Lup93; see problem 18). A lighthouse is positioned at $(x, y) = (\mu, \gamma)$ and it emits discrete light signals in random directions. The coastline is defined as the $y = 0$ line, and the lighthouse's distance from it is γ. Let us define the angle θ as the angle between the line that connects the lighthouse and the point $(x, y) = (\mu, 0)$, and the direction of a signal. The signals will be detected along the coastline with the positions

$$x = \mu + \gamma \tan(\theta), \tag{5.73}$$

[7] It is often said that it takes a 2σ result to convince a theorist that his theory is correct, a 5σ result to convince an observer that an effect is real, and a 10σ result to convince a theorist that his theory is wrong.

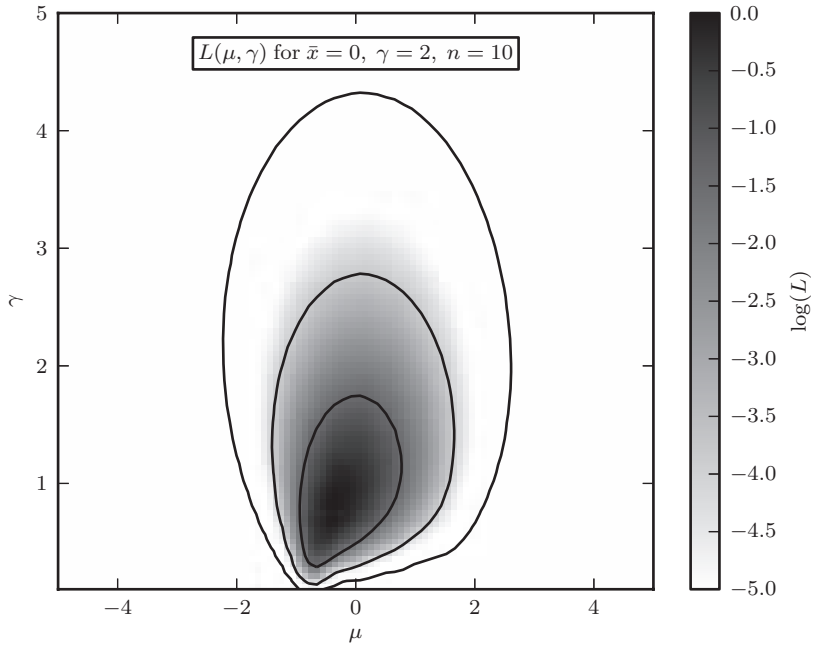

Figure 5.10. An illustration of the logarithm of posterior probability distribution for μ and γ, $L(\mu, \gamma)$ (see eq. 5.75) for $N = 10$ (the sample is generated using the Cauchy distribution with $\mu = 0$ and $\gamma = 2$). The maximum of L is renormalized to 0, and color coded as shown in the legend. The contours enclose the regions that contain 0.683, 0.955 and 0.997 of the cumulative (integrated) posterior probability.

with $-\pi/2 \leq \theta \leq \pi/2$. If the angle θ is distributed uniformly, it is easy to show that x follows the Cauchy distribution given by eq. 3.53 (use $p(x) = (\pi\, dx/d\theta)^{-1} p(\theta)$), and the data likelihood is

$$p(\{x_i\}|\mu, \gamma, I) = \prod_{i=1}^{N} \frac{1}{\pi} \left(\frac{\gamma}{\gamma^2 + (x_i - \mu)^2} \right). \tag{5.74}$$

Given a data set of measured positions $\{x_i\}$, we need to estimate μ and γ. Analogously to the Gaussian case discussed above, we shall adopt a uniform prior distribution for the location parameter μ, and a uniform prior distribution for $\ln \gamma$, for $\mu_{\rm min} \leq \mu \leq \mu_{\rm max}$ and $\gamma_{\rm min} < \gamma \leq \gamma_{\rm max}$. The logarithm of the posterior pdf is

$$L_p \equiv \ln\left[p(\mu, \gamma|\{x_i\}, I)\right] = \text{constant} + (N-1)\, \ln \gamma - \sum_{i=1}^{N} \ln\left[\gamma^2 + (x_i - \mu)^2\right]. \tag{5.75}$$

An example, based on $N = 10$ values of x_i, generated using the Cauchy distribution with $\mu = 0$ and $\gamma = 2$, is shown in figure 5.10. In this particular realization, the maximum of L is at $\mu_0 = -0.36$ and $\gamma_0 = 0.81$. This maximum can be found by setting the derivatives of L to 0 (using eqs. 4.3 and 5.75), but the resulting

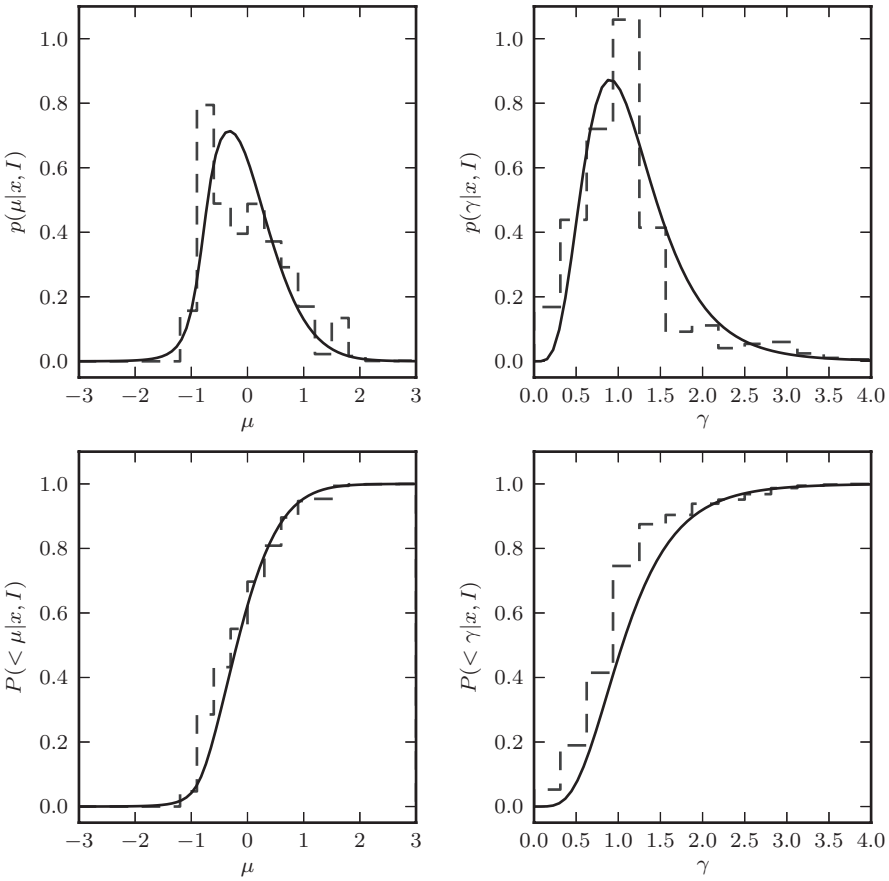

Figure 5.11. The solid lines show the posterior pdf $p(\mu|\{x_i\}, I)$ (top-left panel) and the posterior pdf $p(\gamma|\{x_i\}, I)$ (top-right panel) for the two-dimensional pdf from figure 5.10. The dashed lines show the distribution of approximate estimates of μ and γ based on the median and interquartile range. The bottom panels show the corresponding cumulative distributions.

equations still have to be solved numerically (compare eq. 5.75 to its Gaussian cousin, eq. 5.56; note that in this case we cannot form an analog to eq. 5.56).

The posterior marginal distributions for μ and γ are shown in figure 5.11. Note that the posterior distribution for γ is very asymmetric. For this reason, its peak is not as good an estimator of the true γ as the distribution's median (see the cumulative distribution in the bottom-right panel). As the sample size N increases, both posterior marginal distributions become asymptotically Gaussian and eq. 4.6 can be used to estimate the confidence intervals for μ and γ.

It turns out that for the Cauchy distribution the median and σ_G (see eq. 3.54) can be used as a good shortcut to determine the best-fit parameters (instead of computing marginal posterior pdfs). For example, the sample median and the median of posterior marginal distributions for μ have Kendall's correlation coefficient $\tau \sim 0.7$. For the particular example discussed in figure 5.10, the median and interquartile range imply $\mu = -0.26$ and $\gamma = 1.11$. When using this shortcut, the bootstrap method

can be used to estimate parameter uncertainties (see figure 5.11 for comparison with Bayesian results; the agreement is good though not perfect).

In practice, it is often prohibitive to do an exhaustive search of the parameter space to find the maximum of the posterior pdf (as we did in figures 5.4 and 5.10). Computation becomes especially hard in the case of a high-dimensional parameter space. Instead, various numerical minimization techniques may be used to find the best parameters (discussed in §5.8).

5.6.4. Beating $\sqrt{N}$ for Uniform Distribution

We have discussed how to estimate the location parameter for a uniform distribution in §3.4. Due to the absence of tails, the extreme values of x_i provide a more efficient estimator of the location parameter than the mean value, with errors that improve with the sample size as $1/N$, rather than as $1/\sqrt{N}$ when using the mean. Here we derive this result using the Bayesian method.

Given the uniform distribution described by eq. 3.39, the likelihood of observing x_i is given by

$$p(x_i|\mu, W) = \frac{1}{W} \text{ for } |x_i - \mu| \leq \frac{W}{2}, \tag{5.76}$$

and 0 otherwise (i.e., x_i can be at most $W/2$ away from μ). We assume that both μ and W are unknown and need to be estimated from data. We shall assume a uniform prior for μ between $-\Delta$ and Δ, where Δ greatly exceeds the plausible range for $|\mu + W|$, and a scale-invariant prior for W. The likelihood of observing the observed data set $\{x_i\}$ is

$$p(\mu, W|\{x_i\}, I) \propto \frac{1}{W^{N+1}} \prod_{i=1}^{N} g(\mu, W|x_i), \tag{5.77}$$

where

$$g(\mu, W|x_i) = 1 \text{ for } W \geq 2|x_i - \mu|, \tag{5.78}$$

and 0 otherwise. In the (μ, W) plane, the region allowed by x_i (where $p(\mu, W|\{x_i\}, I) > 0$) is the wedge defined by $W \geq 2(x_i - \mu)$ and $W \geq 2(\mu - x_i)$. For the whole data set, the allowed region becomes a wedge defined by $W \geq 2(x_{\rm max} - \mu)$ and $W \geq 2(\mu - x_{\rm min})$, where $x_{\rm min}$ and $x_{\rm max}$ are the minimum and maximum values in the data set $\{x_i\}$ (see figure 5.12). The minimum allowed W is $W_{\rm min} = (x_{\rm max} - x_{\rm min})$, and the posterior is symmetric with respect to $\tilde{\mu} = (x_{\rm min} + x_{\rm max})/2$. The highest value of the posterior probability is at the point $(\mu = \tilde{\mu}, W = W_{\rm min})$. Within the allowed range, the posterior is proportional to $1/W^{N+1}$ without a dependence on μ, and is 0 otherwise. The logarithm of the posterior probability $p(\mu, W|\{x_i\}, I)$ based on $N = 100$ values of x_i, generated using $\mu = 0$, $W = 1$ (for this particular realization, $x_{\rm min} = 0.047$ and $x_{\rm max} = 9.884$), is shown in figure 5.12. We see that in this case, the likelihood contours are not well described by ellipses!

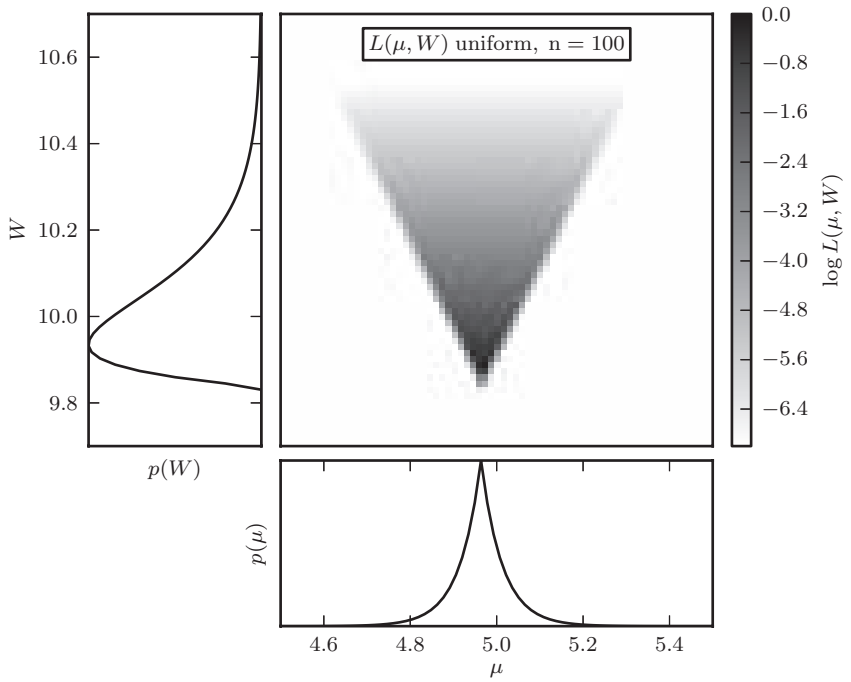

Figure 5.12. An illustration of the logarithm of the posterior probability distribution $L(\mu, W)$ (see eq. 5.77) for $N = 100$, $\mu = 5$, and $W = 10$. The maximum of L is renormalized to 0, and color coded on a scale from -5 to 0, as shown in the legend. The bottom panel shows the marginal posterior for μ (see eq. 5.79), and the left panel shows the marginal posterior for W (see eq. 5.80).

It is straightforward to show that the marginal posterior pdf for μ is

$$p(\mu) = \int_0^\infty p(\mu, W|\{x_i\}, I)\, dW \propto \frac{1}{(|\mu - \tilde{\mu}| + W_{\rm min}/2)^N}, \tag{5.79}$$

and for W,

$$p(W) = \int_{-\Delta}^{\Delta} p(\mu, W|\{x_i\}, I)\, d\mu \propto \frac{W - W_{\rm min}}{W^{N+1}} \text{ for } W \geq W_{\rm min}. \tag{5.80}$$

Taking expectation values for μ and W using these marginal distributions reproduces eqs. 3.68 and 3.69. As shown in figure 5.12, the shape of $p(\mu)$ is more peaked than for a Gaussian distribution.

We can understand how the width of $p(\mu)$ scales with N using eq. 5.79, by considering the scale of the half-max width $d\mu$ satisfying $p(\tilde{\mu} + d\mu) = p(\tilde{\mu})/2$. Putting these values into eq. 5.79 and simplifying leads to

$$d\mu = \frac{W_{\rm min}}{2}\left(2^{1/N} - 1\right). \tag{5.81}$$

Expanding this in an asymptotic series about $N = \infty$ yields the approximation

$$d\mu = \frac{W_{\min} \ln 2}{2N} + \mathcal{O}\left[\left(\frac{1}{N}\right)^2\right], \tag{5.82}$$

which scales as $1/N$, and is in close agreement with the Gaussian approximation to the standard deviation, $(2W_{\min})/(N\sqrt{12})$, derived in §3.4. This $1/N$ scaling is demonstrated using a numerical example shown in figure 3.21.

5.6.5. Parameter Estimation for a Gaussian and a Uniform Background

We shall now solve a slightly more complicated parameter estimation problem involving a Gaussian distribution than those from §5.6.1. Here the underlying model is a mixture of a Gaussian distribution and a uniform distribution in some interval W. This example illustrates parameter covariance and marginalization over a nuisance parameter.

The likelihood of obtaining a measurement x_i is given by

$$p(x_i|A, \mu, \sigma) = \frac{A}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right) + \frac{1 - A}{W} \tag{5.83}$$

for $0 < x < W$, and 0 otherwise. It is implied that the measurement error for x_i is negligible compared to σ (or alternatively that σ is an unknown measurement error when we measure a single-valued quantity μ). We will assume that the location parameter μ for the Gaussian component is known, and that W is sufficiently large so that the Gaussian tails are not truncated. Note that there is a "trade-off" between the first and second component: a larger A means a weaker background component with the strength $B = (1 - A)/W$ (we cannot simply substitute an unknown B for the second term because of the normalization constraint). Since an increase of σ will widen the Gaussian so that its tails partially compensate for the strength of the background component, we expect a covariance between A and σ (because the background strength is not a priori known).

We shall adopt uniform prior distributions for A and σ and require that they are both nonnegative,

$$p(A, \sigma|I) = \text{constant}, \quad \text{for } 0 \le A < A_{\max} \text{ and } 0 \le \sigma < \sigma_{\max}. \tag{5.84}$$

This model might correspond to a spectral line of known central wavelength but with an unknown width (σ) and strength (amplitude A), measured in the presence of a background with strength B. It could also correspond to a profile of a source in an image with an unknown background B; we know the source position and are trying to estimate its flux (A) and width (σ; and perhaps compare it later to the point spread function to check if the source is resolved). Using the numerical methods developed in §5.8, one might even imagine solving a model with all three parameters unknown: location, width, and background level. This extension is addressed in §5.8.6.

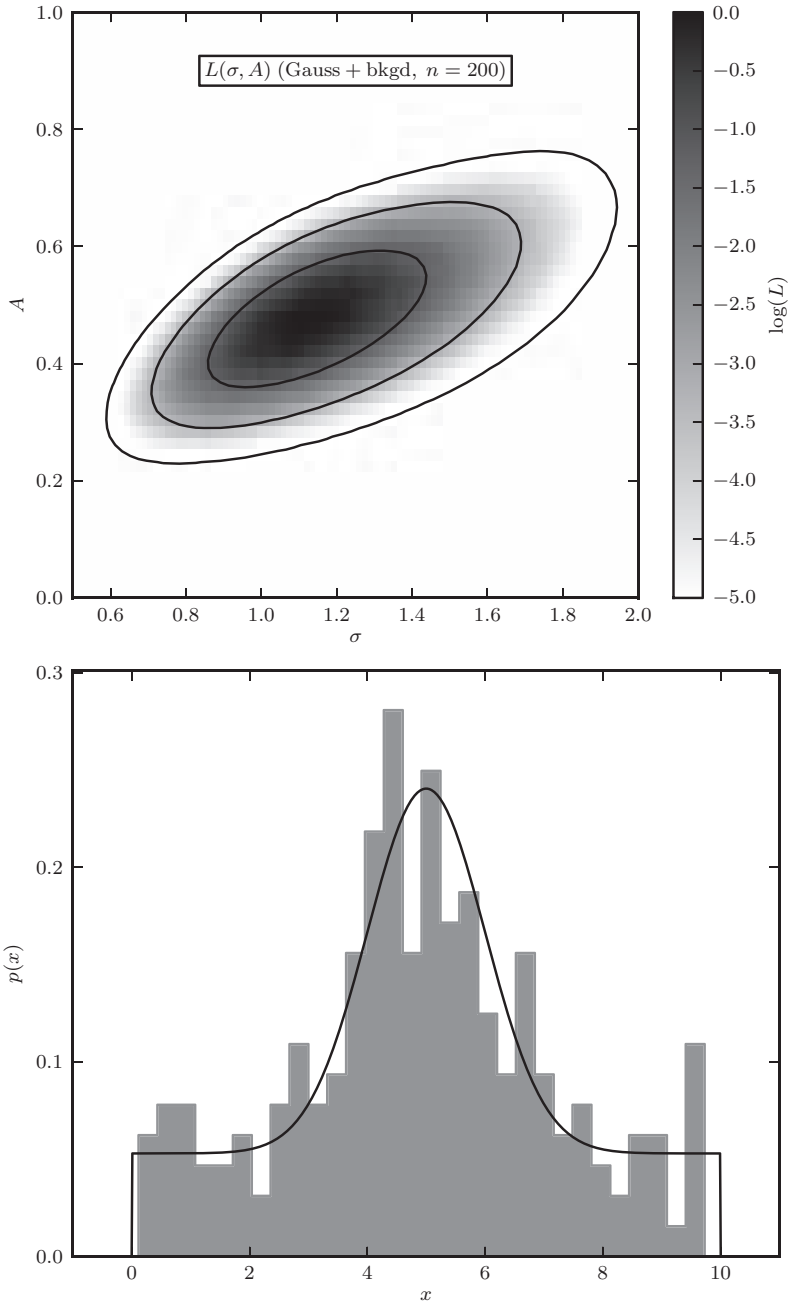

Figure 5.13. An illustration of the logarithm of the posterior probability density function $L(\sigma, A)$ (see eq. 5.85) for data generated using $N = 200$, $\mu = 5$, $\sigma = 1$, and $A = 0.5$, with the background strength $(1 - A)/W = 0.05$ in the interval $0 < x < W$, $W = 10$. The maximum of $L(\sigma, A)$ is renormalized to 0, and color coded on a scale -5 to 0, as shown in the legend. The contours enclose the regions that contain 0.683, 0.955, and 0.997 of the cumulative (integrated) posterior probability. Note the covariance between A and σ. The histogram in the bottom panel shows the distribution of data values used to construct the posterior pdf in the top panel, and the probability density function from which the data were drawn as the solid line.

In the case of a known location μ, the logarithm of the posterior pdf is

$$L_p \equiv \ln\left[p(A, \sigma|\{x_i\}, \mu, W)\right] = \sum_{i=1}^{n} \ln\left[A\frac{\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} + \frac{1-A}{W}\right]. \quad (5.85)$$

It is not possible to analytically find A and σ corresponding to the maximum of this posterior pdf. A numerical example, based on $N = 200$ values of x_i, generated using $A = 0.5$, $\sigma = 1$, $\mu = 5$, and $W = 10$ is shown in figure 5.13.

As expected, there is a trade-off between A and σ: an error in σ is compensated by a proportional error in A. The posterior pdfs for A and σ can be determined by marginalizing the posterior pdf, as we did in §5.6.3 (see figure 5.11). On the other hand, if additional information were available about either parameter, the other one would be better constrained. In this example, if W were significantly increased, the knowledge of the background strength, and thus of A too, would be much improved. In the limit of a perfectly known A, the posterior pdf for σ would simply correspond to a slice through $L_p(A, \sigma)$ (on a linear scale and renormalized!), and would be narrower than the marginal posterior pdf for σ (cf. §3.5.2).

Note that our data $\{x_i\}$ were never binned when computing the posterior probability. We only used the binning and histogram in figure 5.13 to visualize the simulated sample. Nevertheless, the measuring process often results in binned data (e.g., pixels in an image, or spectral measurements). If we had to solve the same problem of a Gaussian line on top of a flat background with binned data, the computation would be similar but not identical. The data would include counts y_i for each bin position x_i. The model prediction for y_i, or $p(D|M, I)$, would still be based on eq. 5.83, and we would need to use counting (Poisson) statistics to describe the variation around the expected count values. This example would be solved similarly to the example discussed next.

5.6.6. Regression Preview: Gaussian vs. Poissonian Likelihood

In examples presented so far, we dealt with distributions of a single random variable. Let us now consider a problem where the data are pairs of random variables, $(x_1, y_1), \ldots, (x_M, y_M)$. Let y correspond to counts in bins centered on x, with a constant and very small bin width. We would like to estimate a and b for a model $y = ax + b$. This is an example of *regression*; this topic is discussed in detail in chapter 8. Here we only want to illustrate how assumptions about data likelihood can affect inferences about model parameters, and discuss the effects of data binning on inferred parameters.

We start the discussion with the unbinned data case, that is, first we will assume that we know the exact x position for each count described by y. We have a data set $\{x_i\}$, of size $N = \sum_{j=1}^{M} y_j$, drawn from a pdf given by $p(x) = ax + b$, with $p(x) = 0$ outside the known range $x_{\min} \leq x \leq x_{\max}$, and we want to estimate a and b. For notational simplicity, we introduce $W = (x_{\max} - x_{\min})$ and $x_{1/2} = (x_{\min} + x_{\max})/2$. It is assumed that the uncertainty for each x_i is negligible compared to W.

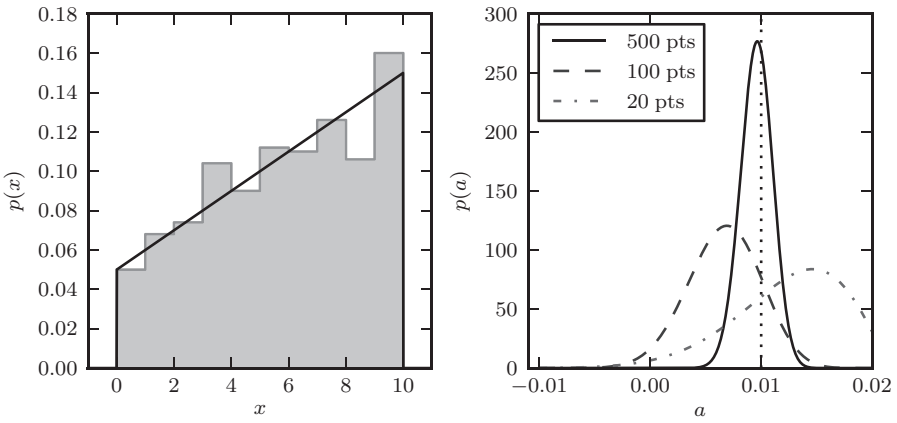

Figure 5.14. Regression of unbinned data. The distribution of $N = 500$ data points is shown in the left panel; the true pdf is shown by the solid curve. Note that although the data are binned in the left panel for visualization purposes, the analysis is performed on the *unbinned* data. The right panel shows the likelihood for the slope a (eq. 5.88) for three different sample sizes. The input value is indicated by the vertical dotted line.

Since $p(x)$ must be a properly normalized pdf, this parameter estimation problem is one-dimensional: a and b are related via the normalization constraint

$$b = \frac{1}{W} - a\, x_{1/2} \tag{5.86}$$

(recall the analysis of truncated data discussed in §4.2.7). The probability of observing a value x_i given a is thus

$$p(x_i|a) = a\,(x_i - x_{1/2}) + \frac{1}{W}. \tag{5.87}$$

Because $p(x) \geq 0$, each data point defines an allowed range for a, $a > -W^{-1}/(x_i - x_{1/2})$. Taking the entire data set, a is confined to the range $a_{\rm min} \leq a \leq a_{\rm max}$, with $a_{\rm min} = -W^{-1}/\max(x_i - x_{1/2})$ and $a_{\rm max} = -W^{-1}/\min(x_i - x_{1/2})$. For $a > 0$, values of x_i larger than $x_{1/2}$ are more probable than smaller ones. Consequently, data values $x_i > x_{1/2}$ favor $a > 0$ and $x_i < x_{1/2}$ favor $a < 0$. Assuming a uniform prior for a, the logarithm of the posterior pdf is

$$L_p(a|\{x_i\}, x_{\rm min}, x_{\rm max}) = \sum_{i=1}^{N} \ln\left[a\,(x_i - x_{1/2}) + \frac{1}{W}\right]. \tag{5.88}$$

For illustration, we generate three data sets with $x_{\rm min} = 0$, $x_{\rm max} = 10$, $a = 0.01$, and, from eq. 5.86, $b = 0.05$. Figure 5.14 shows the distribution of $N = 500$ generated values, and the posterior pdf for a obtained using 20, 100, and all 500 data values.

We now return to the binned data case (recall that the number of data points in the unbinned case is $N = \sum_{i=1}^{M} y_i$). The model for predicting the expected counts

value in a given bin, $y(x) = a^*x + b^*$, can be thought of as

$$y(x) = Y \int_{x-\Delta/2}^{x+\Delta/2} p(x)\, dx, \tag{5.89}$$

where Δ is the bin width and $p(x) = ax + b$ is a properly normalized pdf (i.e., b is related to a via eq. 5.86). The proportionality constant Y provides a scale for the counts and is approximately equal to N (the sum of all counts; we will return to this point below). Here we will assume that Δ is sufficiently small that $p(x)$ is constant within a bin, and thus $a^* = Y\,\Delta\,a$ and $b^* = Y\,\Delta\,b$. When estimating a^* and b^*, Δ need not be explicitly known, as long as it has the same value for all bins and the above assumption that $p(x)$ is constant within a bin holds.

The actual values y_i are "scattered around" their true values $\mu_i = a^*x_i + b^*$, in a manner described by the data likelihood. First we will assume that all μ_i are sufficiently large that the data likelihood can be approximated by a Gaussian, $p(y_i|x_i, a^*, b^*, I) = \mathcal{N}(\mu_i, \sigma_i)$, with $\sigma_i = \mu_i^{1/2}$. With uniform priors for a^* and b^*, the logarithm of the posterior pdf is

$$\begin{aligned} L_p^G(a^*, b^*) \equiv \ln\left[p(a^*, b^*|\{x_i, y_i\})\right] &= \text{constant} \\ &- \frac{1}{2}\sum_{i=1}^{M}\left[\ln(a^*x_i + b^*) + \frac{(y_i - a^*x_i - b^*)^2}{a^*x_i + b^*}\right]. \end{aligned} \tag{5.90}$$

Compared to eq. 5.56, which is also based on a Gaussian likelihood, the term inside the sum is now more complex. Aside from Poisson fluctuations, no source of error is assumed when measuring y_i. In practice, y_i (more precisely, μ_i) can be so large that these counting errors are negligible, and that instead each measurement has an uncertainty σ_i completely unrelated to a^* and b^*. In such a case, the first term in the sum would disappear, and the denominator in the second term would be replaced by σ_i^2.

When the count values are not large, the data likelihood for each y_i must be modeled using the Poisson distribution, $p(k|\mu_i) = \mu_i^k \exp(-\mu_i)/k!$, where $k = y_i$ and as before $\mu_i = a^*x_i + b^*$. With uniform priors for a^* and b^*,

$$L_p^P(a^*, b^*) = \text{constant} + \sum_{i=1}^{M}\left[y_i \ln(a^*x_i + b^*) - a^*x_i - b^*\right]. \tag{5.91}$$

Figure 5.15 compares L_p^G and L_p^P. When the counts in each bin are high, the Gaussian expression L_P^G is a good approximation of the true likelihood L_p^P. When there are fewer counts per bin, the Gaussian approximation becomes biased toward smaller a and b. The difference is not large however: the expectation value and standard deviation for y_i is the same for the Poisson and Gaussian case. It is only the shape of the distribution that changes.

Note also in figure 5.15 that there is a correlation between the best-fit parameters: i.e., the posterior probability contours in the (a^*, b^*) plane are not aligned with the coordinate axes. The orientation of the semimajor axis reflects the normalization

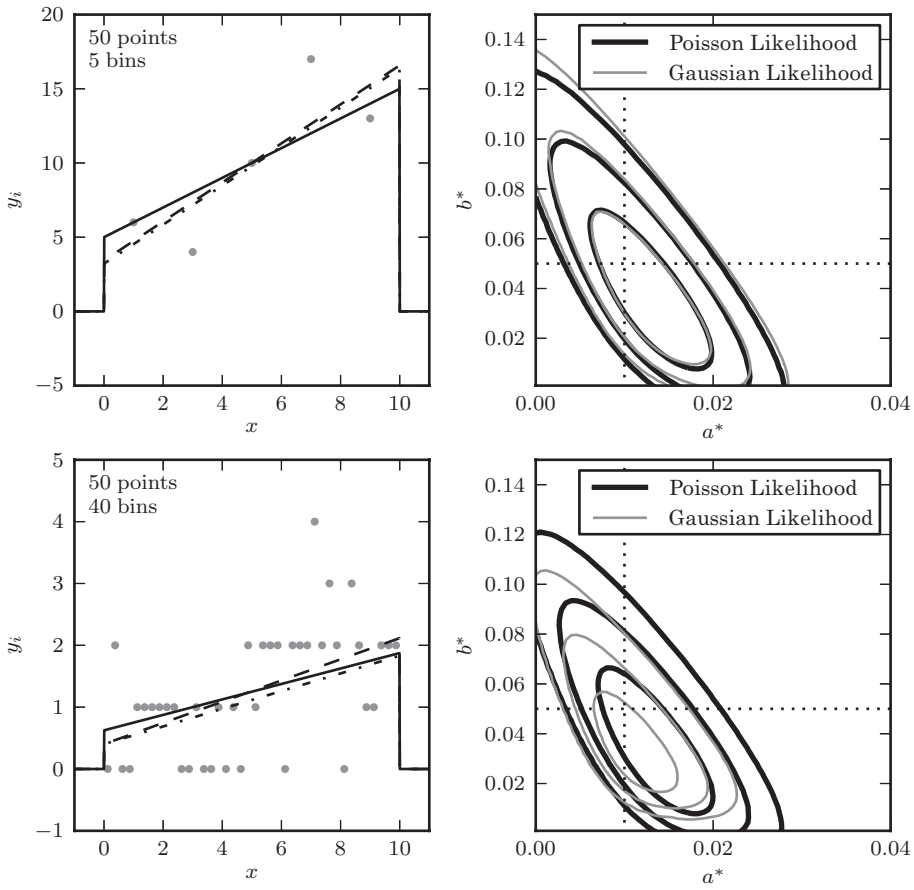

Figure 5.15. Binned regression. The left panels show data sets with 50 points, binned in 5 bins (upper panels) and 40 bins (lower panels). The curves show the input distribution (solid), the Poisson solution (dashed), and the Gaussian solution (dotted). The right panels show 1σ, 2σ, and 3σ likelihood contours for eqs. 5.91 (dark lines) and 5.90 (light lines). With 5 bins (top row) there are enough counts in each bin so that the Gaussian and Poisson predictions are very similar. As the number of bins is increased, the counts decrease and the Gaussian approximation becomes biased.

constraint for the underlying pdf. Analogously to eq. 5.86,

$$b^* = \frac{Y\Delta}{W} - a^* x_{1/2}. \tag{5.92}$$

This equation corresponds to a straight line in the (a^*, b^*) plane that coincides with the major axis of the probability contour for the Poissonian likelihood in figure 5.15. Given this normalization constraint, it may be puzzling that the posterior pdf at a given value of a^* has a finite (nonzero) width in the b^* direction. The reason is that for a given a^* and b^*, the implied Y corresponds to the *predicted* sum of y_i, and the actual sum can fluctuate around the predicted value.

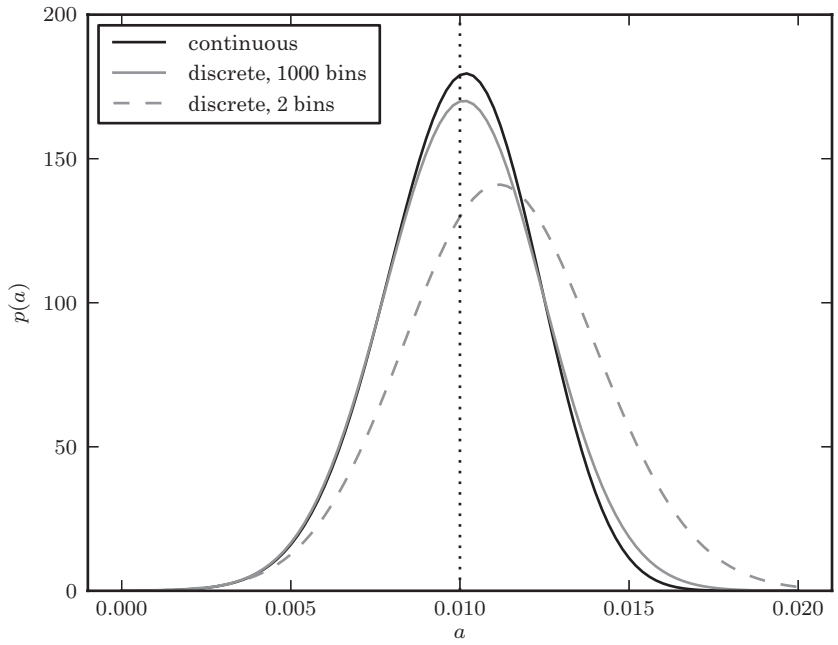

Figure 5.16. The comparison of the continuous method (figure 5.14) and the binned method (figure 5.15) on the same data set. In the limit of a large number of bins, most bins register only zero or one count, and the binned Poisson statistic gives nearly the same marginalized distribution for a as the continuous statistic. For as few as two bins, the constraint on the slope is only slightly biased.

Finally, we would expect that as the number of bins becomes very large, the solution would approach that of the unbinned case. In figure 5.16, we compare the marginalized distribution for a in the unbinned case and the binned case with 1000 bins. The results are nearly identical, as expected.

5.6.7. A Mixture Model: How to "Throw Out" Bad Data Points

We have seen in §5.6.1 and §5.6.3 that standard results for estimating parameters that are based on the assumption of Gaussianity, such as eq. 5.50, do not work well when the sampled distribution is not Gaussian. For example, a few outliers in a data set can significantly affect the weighted mean given by eq. 5.50.

When the model for the underlying distribution is known (e.g., a Cauchy distribution, or a Gaussian convolved with known Gaussian errors), we can maximize the posterior pdf as in any other case. When the model is not known a priori, we can use a Bayesian framework to construct a model in terms of unknown nuisance model parameters, and then marginalize over them to estimate the quantities of interest.

We shall consider here a one-dimensional case and revisit the problem of outliers in more detail in the context of regression (see §8.9). Let us assume as in §5.6.1 that we have a set of N measurements, $\{x_i\}$, of a single-valued quantity μ, and the measurement errors are Gaussian, heteroscedastic, known and given as σ_i. We seek the posterior pdf for μ: $p(\mu|\{x_i\}, \{\sigma_i\})$. The difference compared to §5.6.1 is that here some data points have unreliable σ_i; that is, some points were drawn

from a different distribution with a much larger width. We shall assume no bias for simplicity; that is, these outliers are also drawn from a distribution centered on μ.

If we knew which points were outliers, then we would simply exclude them and apply standard Gaussian results to the remaining points (assuming that outliers represent a small fraction of the data set). We will assume that we do not have this information—this is a case of hidden variables similar to that discussed in §4.4.2, and indeed the analysis will be very similar.

Before proceeding, let us reiterate that an easy and powerful method to estimate μ when we suspect non-Gaussian errors is to simply use the median instead of the mean, and the σ_G instead of the standard deviation in eq. 3.38 when estimating the uncertainty of μ. Nevertheless, the full Bayesian analysis enables a more formal treatment, as well as the ability to estimate which points are likely outliers using an objective framework.

First, given $\{x_i\}$ and $\{\sigma_i\}$, how do we assess whether non-Gaussianity is important? Following the discussion in §4.3.1, we expect that

$$\chi^2_{\rm dof} = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i-\overline{x})^2}{\sigma_i} \approx 1. \tag{5.93}$$

If $\chi^2_{\rm dof} - 1$ is a few times larger than $\sqrt{2/(N-1)}$, then it is unlikely (as given by the cumulative pdf for $\chi^2_{\rm dof}$ distribution) that our data set $\{x_i\}$ was drawn from a Gaussian error distribution with the claimed $\{\sigma_i\}$.

We start by formulating the data likelihood as

$$p(x_i|\mu,\sigma_i,g_i,I) = g_i\,\mathcal{N}(\mu,\sigma_i) + (1-g_i)\,p_{\rm bad}(x_i|\mu,I). \tag{5.94}$$

Here g_i is 1 if the data point is "good" and 0 if it came from the distribution of outliers, $p_{\rm bad}(x_i|\mu,I)$. In this model $p_{\rm bad}(x_i|\mu,I)$ applies to all outliers. Again, if we knew g_i this would be an easy problem to solve. Since $\{g_i\}$ represent hidden variables, we shall treat them as model parameters and then marginalize over them to get $p(\mu|\{x_i\},\{\sigma_i\},I)$. With a separable prior, which implies that the reliability of the measurements is decoupled from the true value of the quantity we are measuring,

$$p(\mu,\{g_i\}|I) = p(\mu|I)\,p(\{g_i\}|I), \tag{5.95}$$

we get

$$p(\mu,\{g_i\}|\{x_i\},I) \propto p(\mu|I)\prod_{i=1}^{N}[g_i\,\mathcal{N}(\mu,\sigma_i) + (1-g_i)\,p_{\rm bad}(x_i|\mu,I)]\;\;p(\{g_i\}|I), \tag{5.96}$$

and finally, marginalizing over g_i gives

$$p(\mu|\{x_i\},\{\sigma_i\}) \propto \int p(\mu,\{g_i\}|\{x_i\},I)\,d^N g_i. \tag{5.97}$$

To proceed further, we need to choose specific models for the priors and $p_{\rm bad}(x_i|\mu,I)$. We shall follow an example from [27], which discusses this problem

in the context of incompatible measurements of the Hubble constant available at the end of the twentieth century.

For the distribution of outliers, we adopt an appropriately wide Gaussian distribution

$$p_{\rm bad}(x_i|\mu, I) = \mathcal{N}(\mu, \sigma^*), \tag{5.98}$$

where a good choice for σ^* is a few times σ_G determined from data, although details vary depending on the actual $\{x_i\}$. If $\sigma^* \gg \sigma_G$, $p_{\rm bad}(x_i|\mu, I)$ effectively acts as a uniform distribution. Alternatively, we could treat σ^* as yet another model parameter and marginalize over it.

The prior for μ can be taken as uniform for simplicity, and there are various possible choices for $p(\{g_i\}|I)$ depending on our beliefs, or additional information, about the quality of measurements. In the example discussing the Hubble constant from [27], each data point is the result of a given method and team, and the prior for g_i might reflect past performance for both. In order to enable an analytic solution here, we shall adopt uniform priors for all g_i. In this case, marginalization over all g_i is effectively replacing every g_i by 1/2, and leads to

$$p(\mu|\{x_i\}, \{\sigma_i\}) \propto \prod_{i=1}^{N} \left[\mathcal{N}(\mu, \sigma_i) + \mathcal{N}(\mu, \sigma^*)\right]. \tag{5.99}$$

Of course, other choices are possible, and they need not be the same for all data points. For example, we could combine uniform priors (no information) with Gaussians, skewed distributions toward $g_i = 0$ or $g_i = 1$, or even delta functions, representing that we are certain that some data points are trustworthy, or more generally, that we know their probability of being correct with certainty.

A distinctive feature of Bayesian analysis is that we can marginalize eq. 5.96 over μ and all g_i but one—we can obtain a posterior probability that a given data point is good (or bad, of course). Assuming uniform priors, let us integrate over all g_i except the first point (without a loss of generality) and μ to get (note that the product starts from $i = 2$)

$$p(\mu, g_1|\{x_i\}, \{\sigma_i\}) = \left[g_1\mathcal{N}(\mu, \sigma_1) + (1-g_1)\mathcal{N}(\mu, \sigma^*)\right] \prod_{i=2}^{N} \left[\mathcal{N}(\mu, \sigma_i) + \mathcal{N}(\mu, \sigma^*)\right]. \tag{5.100}$$

This expression cannot be integrated analytically over μ, but we can easily treat it numerically since it is only two-dimensional. In figure 5.17, we assume a sample of 10 points, with 8 points drawn from $\mathcal{N}(0, 1)$ and 2 points (outliers) drawn from $\mathcal{N}(0, 3)$. We solve eq. 5.100 and plot the likelihood as a function of μ and g_i for two points: a "bad" point and a "good" point. The "bad" point has a maximum at $g_i = 0$: our model identifies it as an outlier. Furthermore, it is clear that as the weight g_i increases, this has a noticeable effect on the value of μ, skewing it toward positive values. The "good" point has a maximum at $g_i = 1$: our model correctly identifies it as coming from the distribution.

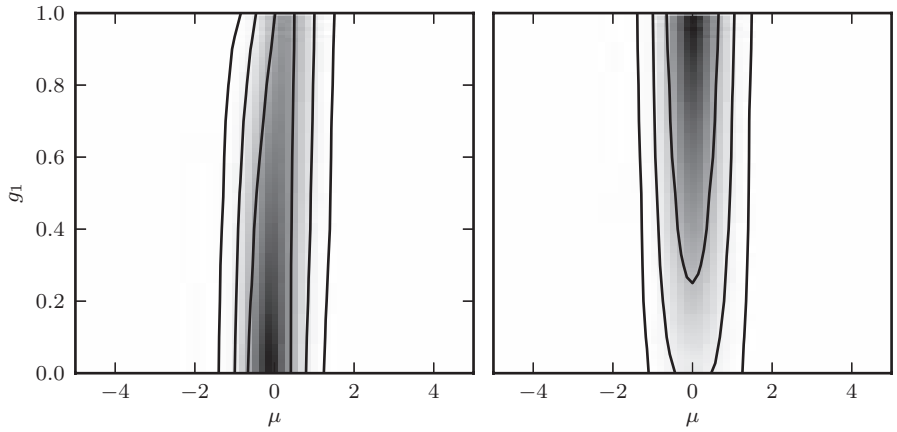

Figure 5.17. The marginal joint distribution between μ and g_i, as given by eq. 5.100. The left panel shows a point identified as bad ($\hat{g}_i = 0$), while the right panel shows a point identified as good ($\hat{g}_i = 1$).

We can also ask about the probability of a point being good or bad, independent of any other parameters. In this case, we would marginalize over μ in eq. 5.100,

$$p(g_1|\{x_i\}, \{\sigma_i\}) = \int p(\mu, g_1|\{x_i\}, \{\sigma_i\})\, d\mu. \tag{5.101}$$

The result is linear in g_i for our uniform priors, and gives the posterior probability of the believability of each point.

We can take an approximate shortcut and evaluate $p(\mu, g_1|\{x_i\}, \{\sigma_i\})$ at the MAP estimate of μ, μ^0, implied by eq. 5.99,

$$p(g_1|\mu^0, \{x_i\}, \{\sigma_i\}) = 2\,\frac{g_1\mathcal{N}(\mu^0, \sigma_1) + (1 - g_1)\mathcal{N}(\mu^0, \sigma^*)}{\left[\mathcal{N}(\mu^0, \sigma_1) + \mathcal{N}(\mu^0, \sigma^*)\right]}, \tag{5.102}$$

where the term in the denominator comes from normalization. With $\sigma^*/\sigma_1 = 3$, it is easy to show that for a likely good point with $x_1 = \mu^0$, $p(g_1) = (1/2) + g_1$, and for a likely bad point a few σ_1 away from μ^0, $p(g_1) = 2(1 - g_1)$. That is, for good data points, the posterior for g_1 is skewed toward 1, and for bad data points it is skewed toward 0. The posterior for g_1 is flat if x_1 satisfies

$$|x_1 - \mu| = \sigma^* \left[\frac{2\ln(\sigma^*/\sigma_1)}{(\sigma^*/\sigma_1)^2 - 1}\right]^{1/2}. \tag{5.103}$$

Although insightful, these results are only approximate. In general, $p(g_1|\mu^0, \{x_i\}, \{\sigma_i\})$ and $p(g_1|\{x_i\}, \{\sigma_i\})$ are not equal, as illustrated in figure 5.18. Note that in the case of the bad point, this approximation leads to a higher likelihood of $g_i = 0$, because moving g_i toward 1 is no longer accompanied by a compensating shift in μ.

We conclude by noting that the data likelihood given by eq. 5.94 is very similar to that encountered in the context of the expectation maximization algorithm

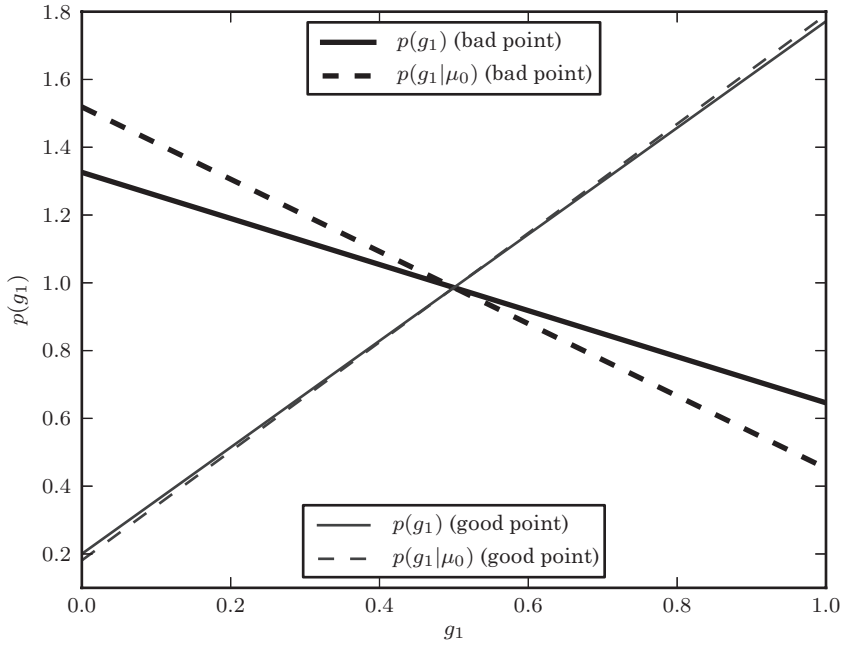

Figure 5.18. The marginal probability for g_i for the "good" and "bad" points shown in figure 5.17. The solid curves show the marginalized probability: that is, eq. 5.100 is integrated over μ. The dashed curves show the probability conditioned on $\mu = \mu_0$, the MAP estimate of μ (eq. 5.102).

(see §4.4.3 and eq. 4.18)—in both cases it is assumed that the data are drawn from a mixture of Gaussians with unknown class labels. If uniform priors are acceptable in a given problem, as we assumed above, then the ideas behind the EM algorithm could be used to efficiently find MAP estimates for both μ and all g_i, without the need to marginalize over other parameters.

5.7. Simple Examples of Bayesian Analysis: Model Selection

5.7.1. Gaussian or Lorentzian Likelihood?

Let us now revisit the examples discussed in §5.6.1 and 5.6.3. In the first example we *assumed* that the data $\{x_i\}$ were drawn from a Gaussian distribution and computed the two-dimensional posterior pdf for its parameters μ and σ. In the second example we did a similar computation, except that we *assumed* a Cauchy (Lorentzian) distribution and estimated the posterior pdf for its parameters μ and γ. What if we do not know what pdf our data were drawn from, and want to find out which of these two possibilities is better supported by our data?

We will assume that the data are identical to those used to compute the posterior pdf for the Cauchy distribution shown in figure 5.10 ($N = 10$, with $\mu = 0$ and $\gamma = 2$). We can integrate the product of the data likelihood and the prior pdf for the

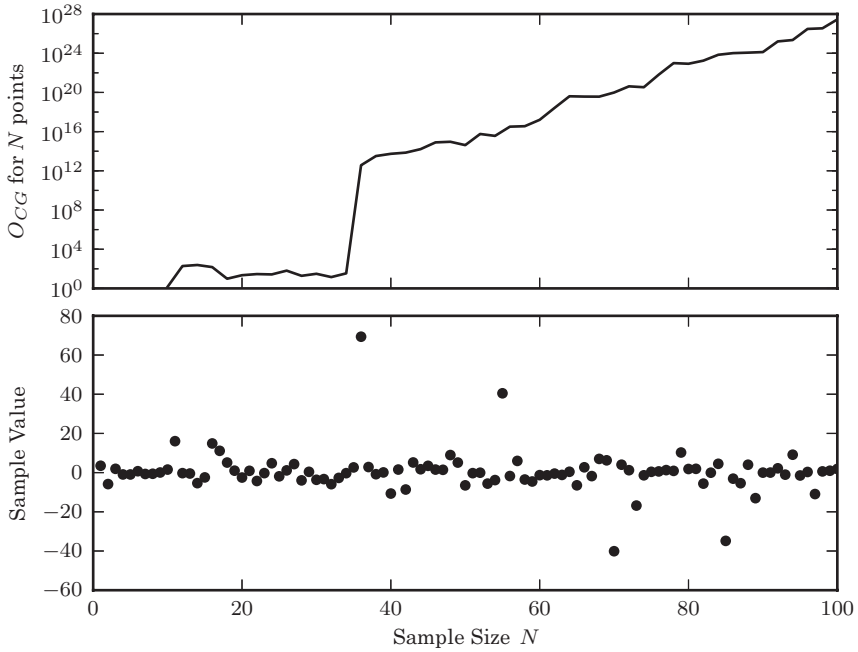

Figure 5.19. The Cauchy vs. Gaussian model odds ratio for a data set drawn from a Cauchy distribution ($\mu = 0$, $\gamma = 2$) as a function of the number of points used to perform the calculation. Note the sharp increase in the odds ratio when points falling far from the mean are added.

model parameters (see eqs. 5.74 and 5.75) to obtain model evidence (see eq. 5.23)

$$E(M = \text{Cauchy}) = \int p(\{x_i\}|\mu, \gamma, I)\, p(\mu, \gamma|I)\, d\mu\, d\gamma = 1.18 \times 10^{-12}. \quad (5.104)$$

When using the pdf illustrated in figure 5.10, we first compute exp(pixel value) for each pixel since the logarithm of the posterior is shown in the figure, then we multiply the result by the pixel area, and then sum all the values. In addition, we need to explicitly evaluate the constant of proportionality (see eq. 5.55). Since we assumed the same priors for both the Gaussian and the Cauchy case, they happen to be irrelevant in this example of a model comparison (but are nevertheless explicitly computed in the code accompanying figure 5.19).

We can construct the posterior pdf for the same data set using the Gaussian posterior pdf given by eq. 5.56 (and explicitly accounting for the proportionality constant) and obtain

$$E(M = \text{Gaussian}) = \int p(\{x_i\}|\mu, \sigma, I)\, p(\mu, \sigma|I)\, d\mu\, d\sigma = 8.09 \times 10^{-13}. \quad (5.105)$$

As no other information is available to prefer one model over the other one, we can assume that the ratio of model priors $p(M_C|I)/p(M_G|I) = 1$, and thus the

odds ratio for the Cauchy vs. Gaussian model is the same as the Bayes factor,

$$O_{\rm CG} = \frac{1.18 \times 10^{-12}}{9.09 \times 10^{-13}} = 1.45. \tag{5.106}$$

The odds ratio is very close to unity and is therefore inconclusive.

Why do we get an inconclusive odds ratio? Recall that this example used a sample of only 10 points; the probability of drawing at least one point far away from the mean, which would strongly argue against the Gaussian model, is fairly small. As the number of data values is increased, the ability to discriminate between the models will increase, too. Figure 5.19 shows the odds ratio for this problem as a function of the number of data points. As expected, when we increase the size of the observed sample, the odds ratio quickly favors the Cauchy over the Gaussian model.

Note the particularly striking feature that the addition of the 36th point causes the odds ratio to jump by many orders of magnitude: this point is extremely far from the mean, and thus is very unlikely under the assumption of a Gaussian model. The effect of this single point on the odds ratio illustrates another important caveat: the presence of even a single outlier may have a large effect on the computed likelihood, and as a result affect the conclusions. If your data has potential outliers, it is very important that these be accounted for within the distribution used for modeling the data likelihood (as was done in §5.6.7).

5.7.2. Understanding Knuth's Histograms

With the material covered in this chapter, we can now return to the discussion of histograms (see §4.8.1) and revisit them from the Bayesian perspective. We pointed out that Scott's rule and the Freedman–Diaconis rule for estimating optimal bin width produce the same answer for multimodal and unimodal distributions as long as their data set size and scale parameter are the same. This undesired result is avoided when using a method developed by Knuth [19]; an earlier discussion of essentially the same method is given in [13].

Knuth shows that the best piecewise constant model has the number of bins, M, which maximizes the following function (up to an additive constant, this is the logarithm of the posterior probability):

$$\begin{aligned} F(M|\{x_i\}, I)) = N \log M + \log\left[\Gamma\left(\frac{M}{2}\right)\right] - M\log\left[\Gamma\left(\frac{1}{2}\right)\right] \\ - \log\left[\Gamma\left(N + \frac{M}{2}\right)\right] + \sum_{k=1}^{M} \log\left[\Gamma\left(n_k + \frac{1}{2}\right)\right], \end{aligned} \tag{5.107}$$

where Γ is the gamma function, and n_k is the number of measurements x_i, $i = 1, \ldots, N$, which are found in bin k, $k = 1, \ldots, M$. Although this expression is more involved than the "rules of thumb" listed in §4.8.1, it can be easily evaluated for an *arbitrary* data set.

Knuth derived eq. 5.107 using Bayesian model selection and treating the histogram as a piecewise constant model of the underlying density function. By assumption, the bin width is constant and the number of bins is the result of

model selection. Given the number of bins, M, the model for the underlying pdf is

$$h(x) = \sum_{k=1}^{M} h_k \Pi(x|x_{k-1}, x_k), \tag{5.108}$$

where the boxcar function $\Pi = 1$ if $x_{k-1} < x \leq x_k$, and 0 otherwise. The M model parameters, $h_k, k = 1, \ldots, M$, are subject to normalization constraints, so that there are only $M-1$ free parameters. The uninformative prior distribution for $\{h_k\}$ is given by

$$p(\{h_k\}|M, I) = \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \left[h_1 h_2 \ldots h_{M-1} \left(1 - \sum_{k=1}^{M-1} h_k \right) \right]^{-1/2}, \tag{5.109}$$

which is known as the Jeffreys prior for the multinomial likelihood. The joint data likelihood is a multinomial distribution (see §3.3.3)

$$p(\{x_i\}|\{h_k\}, M, I) \propto h_1^{n_1} h_2^{n_2} \ldots h_M^{n_M}. \tag{5.110}$$

The posterior pdf for model parameters h_k is obtained by multiplying the prior and data likelihood. The posterior probability for the number of bins M is obtained by marginalizing the posterior pdf over all h_k. This marginalization includes a series of nested integrals over the $(M-1)$-dimensional parameter space, and yields eq. 5.107; details can be found in Knuth's paper.

Knuth also derived the posterior pdf for h_k, and summarized it by deriving its expectation value and variance. The expectation value is

$$h_k = \frac{n_k + \frac{1}{2}}{N + \frac{M}{2}}, \tag{5.111}$$

which is an interesting result (the naive expectation is $h_k = n_k/N$): even when there are no counts in a given bin, $n_k = 0$, we still get a nonvanishing estimate $h_k = 1/(2N + M)$. The reason is that the assumed prior distribution effectively places one half of a datum in each bin.

Comparison of different rules for optimal histogram bin width

The number of bins in Knuth's expression (eq. 5.107) is defined over the observed data range (i.e., the difference between the maximum and minimum value). Since the observed range generally increases with the sample size, it is not obvious how the optimal bin width varies with it. The variation depends on the actual underlying distribution from which data are drawn, and for a Gaussian distribution numerical simulations with N up to 10^6 show that

$$\Delta_b = \frac{2.7\sigma_G}{N^{1/4}}. \tag{5.112}$$

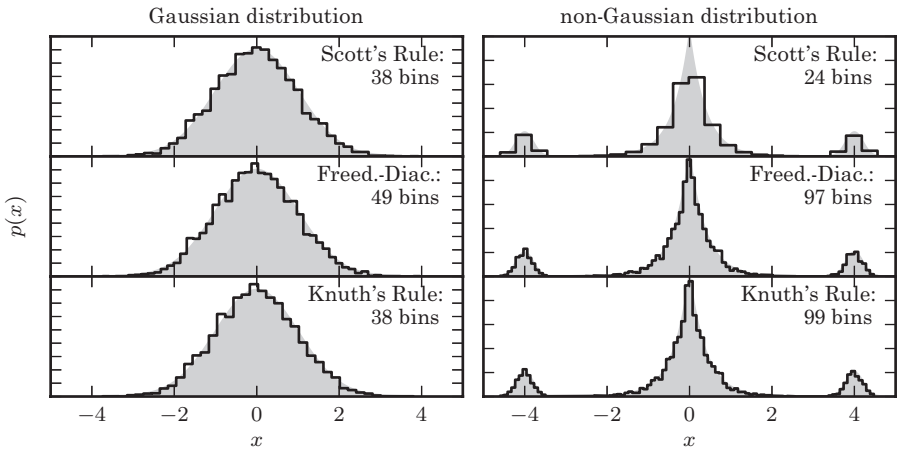

Figure 5.20. The results of Scott's rule, the Freedman–Diaconis rule, and Knuth's rule for selecting the optimal bin width for a histogram. These histograms are based on 5000 points drawn from the shown pdfs. On the left is a simple normal distribution. On the right is a Laplacian distribution at the center, with two small Gaussian peaks added in the wings.

We have deliberately replaced σ by σ_G (see eq. 3.36) to emphasize that the result is applicable to non-Gaussian distributions if they do not show complex structure, such as multiple modes, or extended tails. Of course, for a multimodal distribution the optimal bin width is smaller than given by eq. 5.112 (so that it can "resolve" the substructure in $f(x)$), and can be evaluated using eq. 5.107. Compared to the Freedman–Diaconis rule, the "rule" given by eq. 5.112 has a slower decrease of Δ_b with N; for example, for $N = 10^6$ the Freedman–Diaconis Δ_b is 3 times smaller than that given by eq. 5.112.

Despite the attractive simplicity of eq. 5.112, to utilize the full power of Knuth's method, eq. 5.107 should be used, as done in the following example. Figure 5.20 compares the optimal histogram bins for two different distributions, as selected by Scott's rule, the Freedman–Diaconis rule, and Knuth's method. For the non-Gaussian distribution, Scott's rule greatly underestimates the optimal number of histogram bins, resulting in a histogram that does not give as much intuition as to the shape of the underlying distribution.

The usefulness of Knuth's analysis and the result summarized by eq. 5.107 goes beyond finding the optimal bin size. The method is capable of recognizing substructure in data and, for example, it results in $M = 1$ when the data are consistent with a uniform distribution, and suggests more bins for a multimodal distribution than for a unimodal distribution even when both samples have the same size and σ_G (again, eq. 5.112 is an approximation valid only for unimodal centrally concentrated distributions; if in doubt, use eq. 5.107; for the latter, see the Python code used to generate figure 5.20).

Lastly, remember that Knuth's derivation assumed that the uncertainty of each x_i is negligible. When this is not the case, including the case of heteroscedastic errors, techniques introduced in this chapter can be used for general model selection, including the case of a piecewise constant model, as well as varying bin size.

Bayesian blocks

Though Knuth's Bayesian method is an improvement over the rules of thumb from §4.8.1, it still has a distinct weakness: it assumes a uniform width for the optimal histogram bins. The Bayesian model used to derive Knuth's rule suggests that this limitation could be lifted, by maximizing a well-designed likelihood function over bins of varying width. This approach has been explored in [30, 31], and dubbed *Bayesian blocks*. The method was first developed in the field of time-domain analysis (see §10.3.5), but is readily applicable to histogram data as well; the same ideas are also discussed in [13].

In the Bayesian blocks formalism, the data are segmented into *blocks*, with the borders between two blocks being set by *changepoints*. Using a Bayesian analysis based on Poissonian statistics within each block, an objective function, called the log-likelihood fitness function, can be defined for each block:

$$F(N_i, T_i) = N_i(\log N_i - \log T_i), \tag{5.113}$$

where N_i is the number of points in block i, and T_i is the width of block i (or the duration, in time-series analysis). Because of the additive nature of log-likelihoods, the fitness function for any set of blocks is simply the sum of the fitness functions for each individual block. This feature allows for the configuration space to be explored quickly using dynamic programming concepts: for more information see [31] or the Bayesian blocks implementation in AstroML.

In figure 5.21, we compare a Bayesian blocks segmentation of a data set to a segmentation using Knuth's rule. The adaptive bin width of the Bayesian blocks histogram leads to a better representation of the underlying data, especially when there are fewer points in the data set. An important feature of this method is that the bins are optimal in a quantitative sense, meaning that statistical significance can be attached to the bin configuration. This has led to applications in the field of time-domain astronomy, especially in signal detection.

Finally, we should mention that the fitness function in eq. 5.113 is just one of many possible fitness functions that can be used in the Bayesian blocks method. For more information, see [31] and references therein.

AstroML includes tools for easy computation of the optimal bins derived using Bayesian blocks. The interface is similar to that described in §4.8.1:

```
In [1]: %pylab
In [2]: from astroML.plotting import hist
In [3]: x = np.random.normal(size=1000)
In [4]: hist(x, bins='blocks')  # can also choose
        # bins='knuth'
```

This will internally call the `bayesian_blocks` function in the `astroML.density_estimation` module, and display the resulting histogram. The `hist` function in AstroML operates analogously to the `hist` function in Matplotlib, but can optionally use Bayesian blocks or Knuth's method to choose the binning. For more details see the source code associated with figure 5.21.

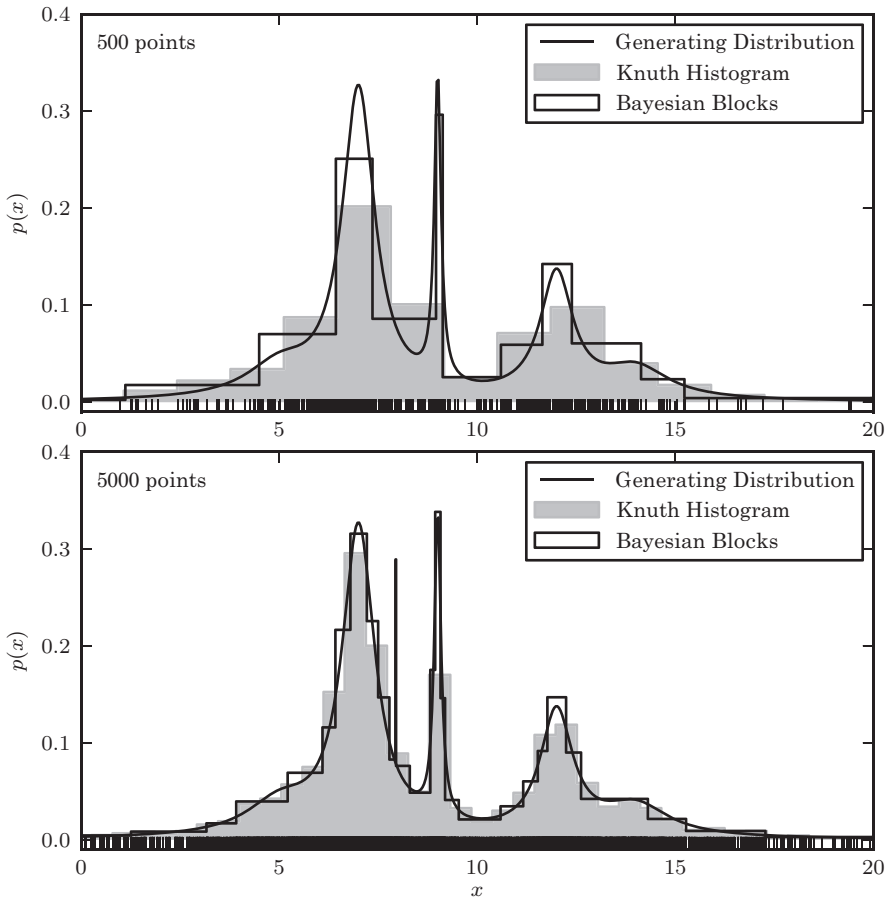

Figure 5.21. Comparison of Knuth's histogram and a Bayesian blocks histogram. The adaptive bin widths of the Bayesian blocks histogram yield a better representation of the underlying data, especially with fewer points.

5.7.3. One Gaussian or Two Gaussians?

In analogy with the example discussed in §5.7.1, we can ask whether our data were drawn from a Gaussian distribution, or from a distribution that can be described as the sum of two Gaussian distributions. In this case, the number of parameters for the two competing models is different: two for a single Gaussian, and five for the sum of two Gaussians. This five-dimensional pdf is hard to treat analytically, and we need to resort to numerical techniques as described in the next section. After introducing these techniques, we will return to this model comparison problem (see §5.8.4).

5.8. Numerical Methods for Complex Problems (MCMC)

When the number of parameters, k, in a model, $M(\boldsymbol{\theta})$, with the vector of parameters $\boldsymbol{\theta}$ specified by θ_p, $p = 1, \ldots, k$, is large, direct exploration of the posterior pdf by

exhaustive search becomes impractical, and often impossible. For example, if the grid for computing the posterior pdf, such as those illustrated in figures 5.4 and 5.10, includes only 100 points per coordinate, the five-dimensional model from the previous example (§5.7.3) will require on order 10^{10} computations of the posterior pdf. Fortunately, a number of numerical methods exist that utilize more efficient approaches than an exhaustive grid search.

Let us assume that we know how to compute the posterior pdf (we suppress the vector notation for $\boldsymbol{\theta}$ for notational clarity since in the rest of this section we always discuss multidimensional cases)

$$p(\theta) \equiv p(M(\boldsymbol{\theta})|D, I) \propto p(D|M(\boldsymbol{\theta}), I)p(\boldsymbol{\theta}|I). \tag{5.114}$$

In general, we wish to evaluate the multidimensional integral

$$I(\theta) = \int g(\theta)p(\theta)\, d\theta. \tag{5.115}$$

There are two classes of frequently encountered problems:

1. *Marginalization and parameter estimation*, where we seek the posterior pdf for parameters θ_i, $i = 1, \ldots, P$, and the integral is performed over the space spanned by nuisance parameters θ_j, $j = (P + 1), \ldots, k$ (for notational simplicity we assume that the last $k - P$ parameters are nuisance parameters). In this case, $g(\theta) = 1$. As a special case, we can seek the posterior mean (see eq. 5.7) for parameter θ_m, where $g(\theta) = \theta_m$, and the integral is performed over all other parameters. Analogously, we can also compute the credible region, defined as the interval that encloses $1 - \alpha$ of the posterior probability. In all of these computations, it is sufficient to evaluate the integral in eq. 5.115 up to an unknown normalization constant because the posterior pdf can be renormalized to integrate to unity.
2. *Model comparison*, where $g(\theta) = 1$ and the integral is performed over all parameters (see eq. 5.23). Unlike the first class of problems, here the proper normalization is mandatory.

One of the simplest numerical integration methods is generic Monte Carlo. We generate a random set of M values θ, θ_j, $j = 1, \ldots, M$, uniformly sampled within the integration volume V_θ, and estimate the integral from eq. 5.115 as

$$I \approx \frac{V_\theta}{M} \sum_{j=1}^{M} g(\theta_j)\, p(\theta_j). \tag{5.116}$$

This method is very inefficient when the integrated function greatly varies within the integration volume, as is the case for the posterior pdf. This problem is especially acute with high-dimensional integrals.

A number of methods exist that are much more efficient than generic Monte Carlo integration. The most popular group of techniques is known as Markov chain Monte Carlo (MCMC) methods. They return a sample of points, or chain, from the k-dimensional parameter space, with a distribution that is asymptotically

proportional to $p(\theta)$. The constant of proportionality is not important in the first class of problems listed above. In model comparison problems, the proportionality constant from eq. 5.117 must be known; we return to this point in §5.8.4.

Given such a chain of length M, the integral from eq. 5.115 can be estimated as

$$I = \frac{1}{M}\sum_{j=1}^{M} g(\theta_j). \tag{5.117}$$

As a simple example, to estimate the expectation value for θ_1 (i.e., $g(\theta) = \theta_1$), we simply take the mean value of all θ_1 in the chain.

Given a Markov chain, quantitative description of the posterior pdf becomes a density estimation problem (density estimation methods are discussed in Chapter 6). To visualize the posterior pdf for parameter θ_1, marginalized over all other parameters, $\theta_2, \ldots, \theta_k$, we can construct a histogram of all θ_1 values in the chain, and normalize its integral to 1. To get a MAP estimate for θ_1, we find the maximum of this marginalized pdf. A generalization of this approach to multidimensional projections of the parameter space is illustrated in figure 5.22.

5.8.1. Markov Chain Monte Carlo

A Markov chain is a sequence of random variables where a given value nontrivially depends *only on its preceding value*. That is, given the *present* value, past and future values are independent. In this sense, a Markov chain is "memoryless." The process generating such a chain is called the Markov process and can be described as

$$p(\theta_{i+1}|\{\theta_i\}) = p(\theta_{i+1}|\theta_i), \tag{5.118}$$

that is, the next value depends only on the current value.

In our context, θ can be thought of as a vector in multidimensional space, and a realization of the chain represents a path through this space. To reach an equilibrium, or stationary, distribution of positions, it is necessary that the transition probability is symmetric:

$$p(\theta_{i+1}|\theta_i) = p(\theta_i|\theta_{i+1}). \tag{5.119}$$

This condition is called the detailed balance or reversibility condition. It shows that the probability of a jump between two points does not depend on the direction of the jump.

There are various algorithms for producing Markov chains that reach some prescribed equilibrium distribution, $p(\theta)$. The use of resulting chains to perform Monte Carlo integration of eq. 5.115 is called *Markov chain Monte Carlo* (MCMC).

5.8.2. MCMC Algorithms

Algorithms for generating Markov chains are numerous and greatly vary in complexity and applicability. Many of the most important ideas were generated in physics, especially in the context of statistical mechanics, thermodynamics, and quantum field theory [23]. We will only discuss in detail the most famous Metropolis–Hastings algorithm, and refer the reader to Greg05 and BayesCosmo, and references therein, for a detailed discussion of other algorithms.

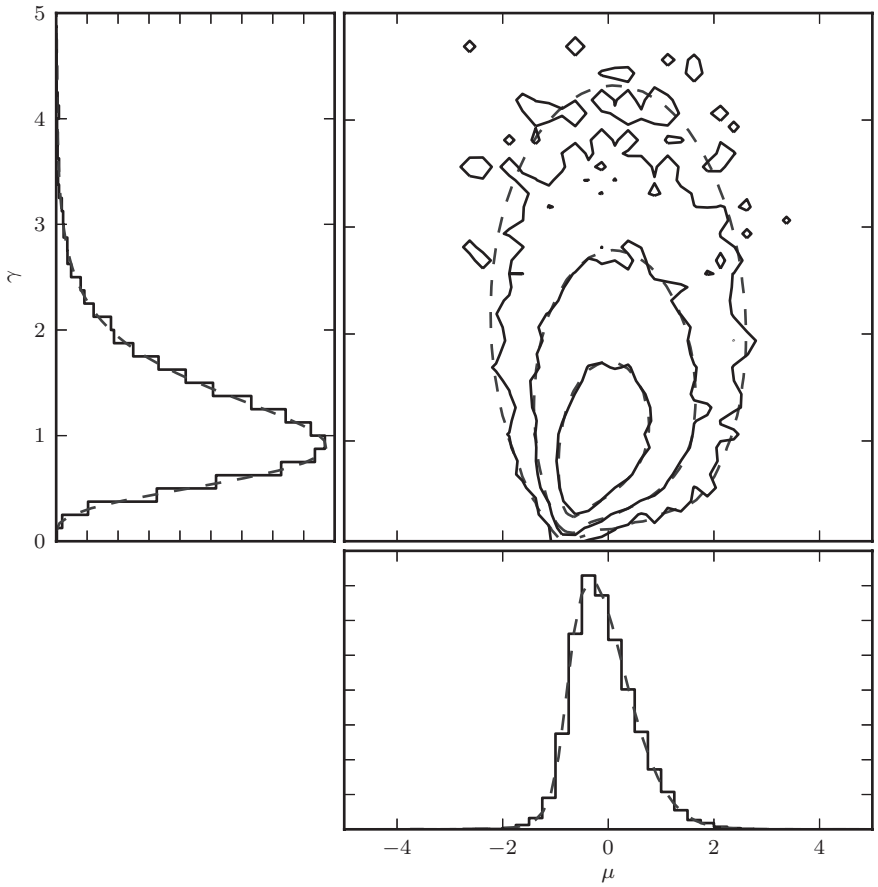

Figure 5.22. Markov chain Monte Carlo (MCMC) estimates of the posterior pdf for parameters describing the Cauchy distribution. The data are the same as those used in figure 5.10: the dashed curves in the top-right panel show the results of direct computation on a regular grid from that diagram. The solid curves are the corresponding MCMC estimates using 10,000 sample points. The left and the bottom panels show marginalized distributions.

In order for a Markov chain to reach a stationary distribution proportional to $p(\theta)$, the probability of arriving at a point θ_{i+1} must be proportional to $p(\theta_{i+1})$,

$$p(\theta_{i+1}) = \int T(\theta_{i+1}|\theta_i)\, p(\theta_i)\, d\theta_i, \tag{5.120}$$

where the transition probability $T(\theta_{i+1}|\theta_i)$ is called the jump kernel or transition kernel (and it is assumed that we know how to compute $p(\theta_i)$). This requirement will be satisfied when the transition probability satisfies the detailed balance condition

$$T(\theta_{i+1}|\theta_i)\, p(\theta_i) = T(\theta_i|\theta_{i+1})\, p(\theta_{i+1}). \tag{5.121}$$

Various MCMC algorithms differ in their choice of transition kernel (see Greg05 for a detailed discussion).

The Metropolis–Hastings algorithm adopts the kernel

$$T(\theta_{i+1}|\theta_i) = p_{\text{acc}}(\theta_i, \theta_{i+1})\, K(\theta_{i+1}|\theta_i), \tag{5.122}$$

where the proposed density distribution $K(\theta_{i+1}|\theta_i,)$ is an *arbitrary* function. The proposed point θ_{i+1} is randomly accepted with the acceptance probability

$$p_{\text{acc}}(\theta_i, \theta_{i+1}) = \frac{K(\theta_i|\theta_{i+1})\, p(\theta_{i+1})}{K(\theta_{i+1}|\theta_i)\, p(\theta_i)} \tag{5.123}$$

(when exceeding 1, the proposed point θ_{i+1} is always accepted). When θ_{i+1} is rejected, θ_i is added to the chain instead. A Gaussian distribution centered on θ_i is often used for $K(\theta_{i+1}|\theta_i)$.

The original Metropolis algorithm is based on a symmetric proposal distribution, $K(\theta_{i+1}|\theta_i) = K(\theta_i|\theta_{i+1})$, which then cancels out from the acceptance probability. In this case, θ_{i+1} is always accepted if $p(\theta_{i+1}) > p(\theta_i)$, and if not, then it is accepted with a probability $p(\theta_{i+1})/p(\theta_i)$.

Although $K(\theta_{i+1}|\theta_i)$ satisfies a Markov chain requirement that it must be a function of only the current position θ_i, it takes a number of steps to reach a stationary distribution from an initial arbitrary position θ_0. These early steps are called the "burn-in" and need to be discarded in analysis. There is no general theory for finding transition from the burn-in phase to the stationary phase; several methods are used in practice. Gelman and Rubin proposed to generate a number of chains and then compare the ratio of the variance between the chains to the mean variance within the chains (this ratio is known as the R statistic). For stationary chains, this ratio will be close to 1. The autocorrelation function (see §10.5) for the chain can be used to determine the required number of evaluations of the posterior pdf to get estimates of posterior quantities with the desired precision; for a detailed practical discussion see [7]. The autocorrelation function can also be used to estimate the increase in Monte Carlo integration error due to the fact that the sequence is correlated (see eq. 10.93).

When the posterior pdf is multimodal, the simple Metropolis–Hastings algorithm can become stuck in a local mode and not find the globally best mode within a reasonable running time. There are a number of better algorithms, such as Gibbs sampling, parallel tempering, various genetic algorithms, and nested sampling. For a good overview, see [3].

5.8.3. PyMC: MCMC in Python

For the MCMC examples in this book, we use the Python package PyMC.[8] PyMC comprises a set of flexible tools for performing MCMC using the Metropolis–Hastings algorithm, as well as maximum a priori estimates, normal approximations, and other sampling techniques. It includes built-in models for common distributions and priors (e.g. Gaussian distribution, Cauchy distribution, etc.) as well as an easy framework to define arbitrarily complicated distributions. For examples of the use of PyMC in practice, see the code accompanying MCMC figures throughout this text.

[8] `https://github.com/pymc-devs/pymc`

While PyMC offers some powerful tools for fine-tuning of MCMC chains, such as varying step methods, fitting algorithms, and convergence diagnostics, for simplicity we use only the basic features for the examples in this book. In particular, the burn-in for each chain is accomplished by simply setting the burn-in size high enough that we can assume the chain has become stationary. For more rigorous approaches to this, as well as details on the wealth of diagnostic tools available, refer to the PyMC documentation.

A simple fit with PyMC can be accomplished as follows. Here we will fit the mean of a distribution—perhaps an overly simplistic example for MCMC, but useful as an introductory example:

```
import numpy as np
import pymc

# generate random Gaussian data with mu=0, sigma=1
N = 100
x = np.random.normal(size=N)

# define the MCMC model: uniform prior on mu,
  # fixed (known) sigma
mu = pymc.Uniform('mu', -5, 5)
sigma = 1
M = pymc.Normal('M', mu, sigma, observed=True,
                value=x)
model = dict(M=M, mu=mu)

# run the model, and get the trace of mu
S = pymc.MCMC(model)
S.sample(10000, burn=1000)
mu_sample = S.trace('mu')[:]

# print the MCMC estimate
print("Bayesian (MCMC): %.3f +/- %.3f"
      % (np.mean(mu_sample), np.std(mu_sample)))

# compare to the frequentist estimate
print("Frequentist: %.3f +/- %.3f"
      % (np.mean(x), np.std(x, ddof=1) / np.sqrt(N)))
```

The resulting output for one particular random seed:

```
Bayesian (MCMC): -0.054 +/- 0.103
Frequentist: -0.050 +/- 0.096
```

As expected for a uniform prior on μ, the Bayesian and frequentist estimates (via eqs. 3.31 and 3.34) are consistent. For examples of higher-dimensional MCMC problems, see the online source code associated with the MCMC figures throughout the text.

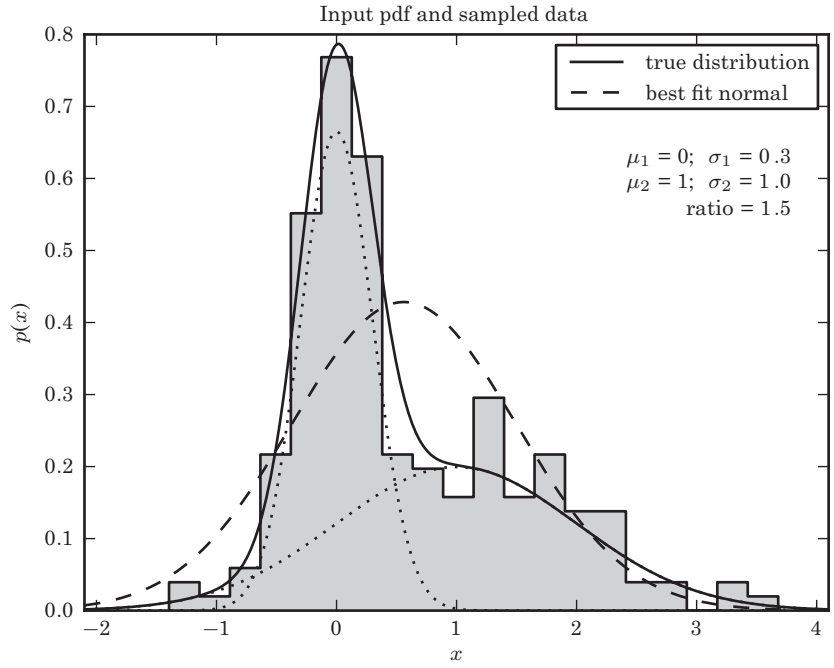

Figure 5.23. A sample of 200 points drawn from a Gaussian mixture model used to illustrate model selection with MCMC.

PyMC is far from the only option for MCMC computation in Python. One other tool that deserves mention is `emcee`,[9] a package developed by astronomers, which implements a variant of MCMC where the sampling is invariant to affine transforms (see [7, 11]). Affine-invariant MCMC is a powerful algorithm and offers improved runtimes for some common classes of problems.

5.8.4. Example: Model Selection with MCMC

Here we return to the problem of model selection from a Bayesian perspective. We have previously mentioned the odds ratio (§5.4), which takes into account the entire posterior distribution, and the Aikake and Bayesian information criteria (AIC and BIC—see §5.4.3), which are based on normality assumptions of the posterior.

Here we will examine an example of distinguishing between unimodal and bimodal models of a distribution in a Bayesian framework. Consider the data sample shown in figure 5.23. The sample is drawn from a bimodal distribution: the sum of two Gaussians, with the parameter values indicated in the figure. The best-fit normal distribution is shown as a dashed line. The question is, can we use a Bayesian framework to determine whether a single-peak or double-peak Gaussian is a better fit to the data?

A double Gaussian model is a five-parameter model: the first four parameters include the mean and width for each distribution, and the fifth parameter is the

[9] Cleverly dubbed "MCMC Hammer," `http://danfm.ca/emcee/`

TABLE 5.2.
Comparison of the odds ratios for a single and double Gaussian model using maximum a posteriori log-likelihood, AIC, and BIC.

	M1: single Gaussian	M2: double Gaussian	M1 − M2
$-2\ln L^0$	465.4	406.0	59.4
BIC	476.0	432.4	43.6
AIC	469.4	415.9	53.5

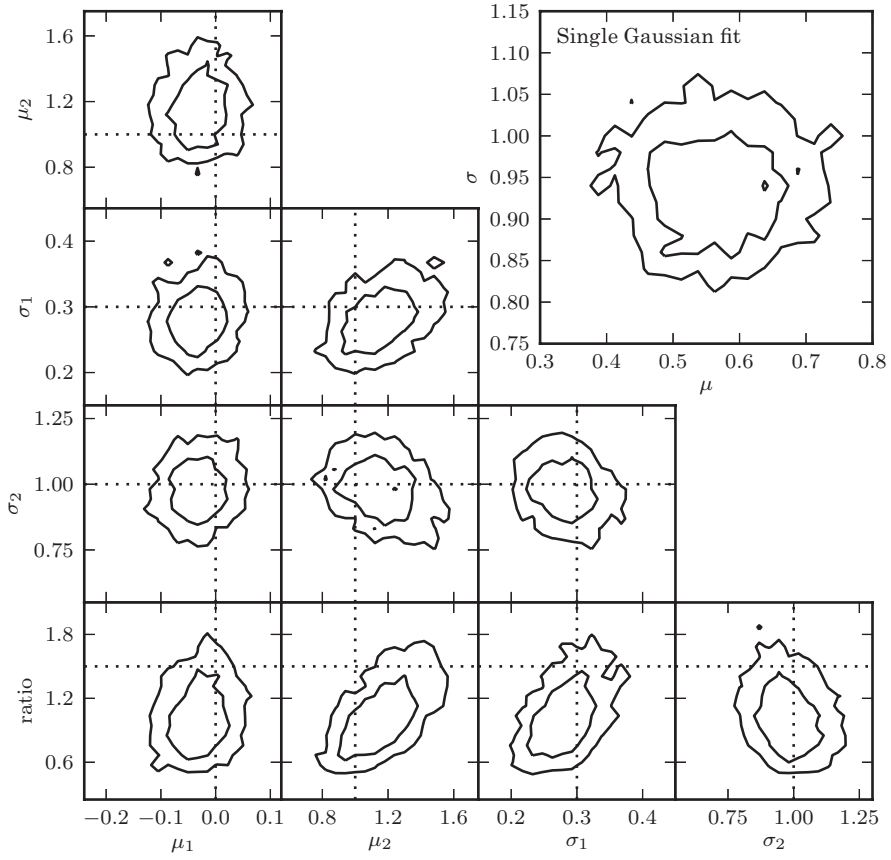

Figure 5.24. The top-right panel shows the posterior pdf for μ and σ for a single Gaussian fit to the data shown in figure 5.23. The remaining panels show the projections of the five-dimensional pdf for a Gaussian mixture model with two components. Contours are based on a 10,000 point MCMC chain.

relative normalization (weight) of the two components. Computing the AIC and BIC for the two models is relatively straightforward: the results are given in table 5.2, along with the maximum a posteriori log-likelihood $\ln L^0$ (the code for maximization of the likelihood and computation of the AIC/BIC can be found in the source of figure 5.24).

It is clear that by all three measures, the double Gaussian model is preferred. But these measures are only accurate if the posterior distribution is approximately Gaussian. For non-Gaussian posteriors, the best statistic to use is the odds ratio (§5.4). While odds ratios involving two-dimensional posteriors can be computed relatively easily (see §5.7.1), integrating five-dimensional posteriors is computationally difficult. This is one manifestation of the curse of dimensionality (see §7.1). So how do we proceed? One way to estimate an odds ratio is based on MCMC sampling.

Computing the odds ratio involves integrating the unnormalized posterior for a model (see §5.7.1):

$$L(M) = \int p(\boldsymbol{\theta}|\{x_i\}, I)\, d^k\theta, \tag{5.124}$$

where the integration is over all k model parameters. How can we compute this based on an MCMC sample? Recall that the set of points derived by MCMC is designed to be distributed according to the posterior distribution $p(\boldsymbol{\theta}|\{x_i\}, I)$, which we abbreviate to simply $p(\boldsymbol{\theta})$. This means that the local density of points $\rho(\boldsymbol{\theta})$ is proportional to this posterior distribution: for a well-behaved MCMC chain with N points,

$$\rho(\boldsymbol{\theta}) = C\, N\, p(\boldsymbol{\theta}), \tag{5.125}$$

where C is an unknown constant of proportionality. Integrating both sides of this equation and using $\int \rho(\boldsymbol{\theta})\, d^k\theta = N$, we find

$$L(M) = 1/C. \tag{5.126}$$

This means that at each point $\boldsymbol{\theta}$ in parameter space, we can estimate the integrated posterior using

$$L(M) = \frac{Np(\boldsymbol{\theta})}{\rho(\boldsymbol{\theta})}. \tag{5.127}$$

We see that the result can be computed from quantities that can be estimated from the MCMC chain: $p(\boldsymbol{\theta}_i)$ is the posterior evaluation at each point, and the local density $\rho(\boldsymbol{\theta}_i)$ can be estimated from the local distribution of points in the chain. The odds ratio problem has now been expressed as a density estimation problem, which can be approached in a variety of ways; see [3, 12]. Several relevant tools and techniques can be found in chapter 6. Because we can estimate the density at the location of each of the N points in the MCMC chain, we have N separate estimators of $L(M)$.

Using this approach, we can evaluate the odds ratio for model 1 (a single Gaussian: 2 parameters) vs. model 2 (two Gaussians: 5 parameters) for our example data set. Figure 5.24 shows the MCMC-derived likelihood contours (using 10,000 points) for each parameter in the two models. For model 1, the contours appear to be nearly Gaussian. For model 2, they are further from Gaussian, so the AIC and BIC values become suspect.

Using the density estimation procedure above,[10] we compute the odds ratio $O_{21} \equiv L(M_2)/L(M_1)$ and find that $O_{21} \approx 10^{11}$, strongly in favor of the

[10]We use a *kernel density estimator* here, with a top-hat kernel for computational simplicity; see §6.1.1 for details.

two-peak solution. For comparison, the implied difference in BIC is $-2\ln(O_{21}) = 50.7$, compared to the approximate value of 43.6 from table 5.2. The Python code that implements this estimation can be found in the source of figure 5.24.

5.8.5. Example: Gaussian Distribution with Unknown Gaussian Errors

In §5.6.1, we explored several methods to estimate parameters for a Gaussian distribution from data with heteroscedastic errors e_i. Here we take this to the extreme, and allow each of the errors e_i to vary as part of the model. Thus our model has $N+2$ parameters: the mean μ, the width σ, and the data errors e_i, $i = 1, \ldots, N$. To be explicit, our model here (cf. eq. 5.63) is given by

$$p(\{x_i\}|\mu, \sigma, \{e_i\}, I) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}(\sigma^2 + e_i^2)^{1/2}} \exp\left(\frac{-(x_i - \mu)^2}{2(\sigma^2 + e_i^2)}\right). \tag{5.128}$$

Though this pdf cannot be maximized analytically, it is relatively straightforward to compute via MCMC, by setting appropriate priors and marginalizing over the e_i as nuisance parameters. Because the e_i are scale factors like σ, we give them scale-invariant priors.

There is one interesting detail about this choice. Note that because σ and e_i appear together as a sum, the likelihood in eq. 5.128 has a distinct degeneracy. For any point in the model space, an identical likelihood can be found by scaling $\sigma^2 \to \sigma^2 + K, e_i^2 \to e_i^2 - K$ for all i (subject to positivity constraints on each term). Moreover, this degeneracy exists at the maximum just as it does elsewhere. Because of this, using priors of different forms on σ and e_i can lead to suboptimal results. If we chose, for example, a scale-invariant prior on σ and a flat prior on e_i, then our posterior would strongly favor $\sigma \to 0$, with the e_i absorbing its effect. This highlights the importance of carefully choosing priors on model parameters, even when those priors are flat or uninformative!

The result of an MCMC analysis on all $N+2$ parameters, marginalized over e_i, is shown in figure 5.25. For comparison, we also show the contours from figure 5.7. The input distribution is within 1σ of the most likely marginalized result, and this is with *no prior knowledge* about the error in each point!

5.8.6. Example: Unknown Signal with an Unknown Background

In §5.6.5 we explored Bayesian parameter estimation for the width of a Gaussian in the presence of a uniform background. Here we consider a more general model and find the width σ and location μ of a Gaussian signal within a uniform background. The likelihood is given by eq. 5.83, where σ, μ, and A are unknown. The results are shown in figure 5.26. The procedure for fitting this, which can be seen in the online source code for figure 5.26, is very general. If the signal shape were not Gaussian, it would be easy to modify this procedure to include another model. We could also evaluate a range of possible signal shapes and compare the models using the model odds ratio, as we did above.

Note that here the data are unbinned; if the data were binned (i.e., if we were trying to fit the number of counts in a data histogram), then this would be very similar to the matched filter analysis discussed in §10.4.

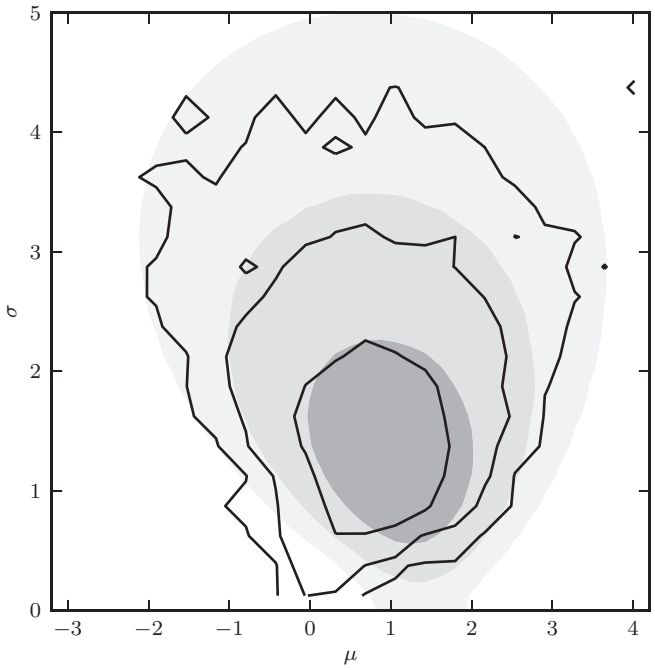

Figure 5.25. The posterior pdf for μ and σ for a Gaussian distribution with heteroscedastic errors. This is the same data set as used in figure 5.7, but here each measurement error is assumed unknown, treated as a model parameter with a scale-invariant prior, and marginalized over to obtain the distribution of μ and σ shown by contours. For comparison, the posterior pdf from figure 5.7 is shown by shaded contours.

5.9. Summary of Pros and Cons for Classical and Bayesian Methods

With the material covered in the last two chapters, we can now compare and contrast the frequentist and Bayesian approaches. It is possible to live completely in either paradigm as a data analyst, performing all of the needed types of analysis tasks that come up in real problems in some manner. So, what should an aspiring young—or not so young—scientist do?

Technical differences We will first discuss differences of a technical nature, then turn to subjective issues. Volumes of highly opinionated text have been written arguing for each of the two sides over many decades (whose main points we can only summarize), with no clear victory yet as of the writing of this book. Given this, the eagerly bellicose partisan looking to eliminate the other side should not be surprised to find that, upon deeper examination, the issues are complex and subtle, and any win is mixed and partial at best. A brief visit to some common battlegrounds, in the form of various basic inference tasks, reveals this:

- **Point estimates**: We can quantify the comparison through an appeal to decision theory, which is the study, using asymptotic theory, of the relative quality of estimators as the amount of data increases. What do we find when

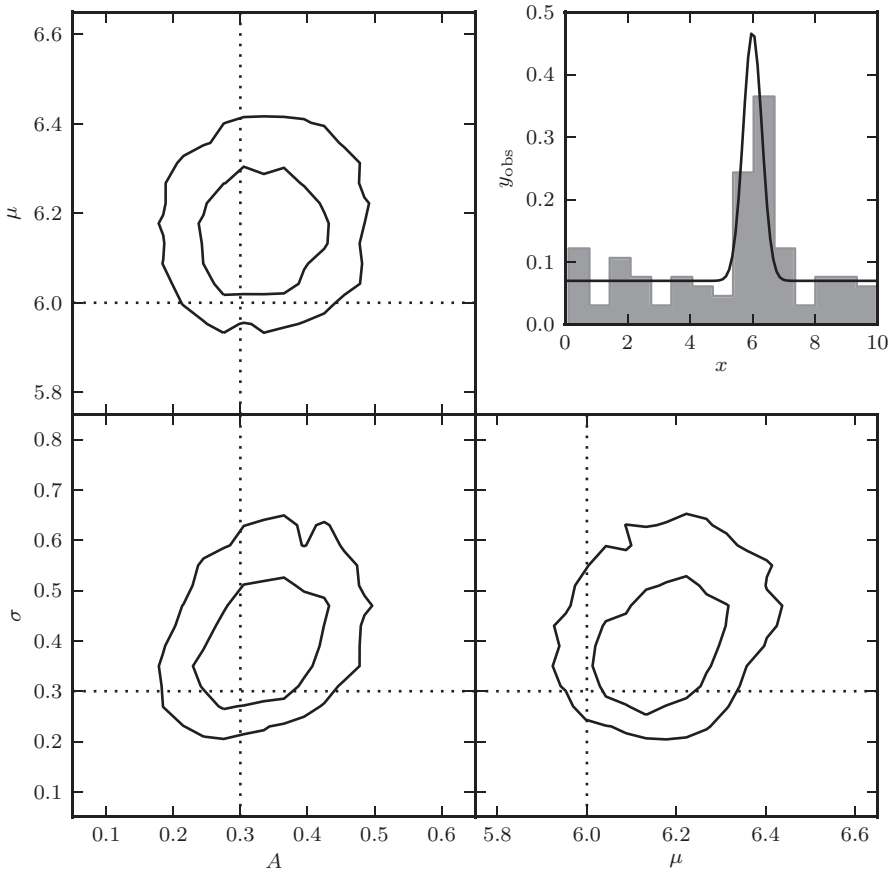

Figure 5.26. Fitting a model of a signal in an unknown background. The histogram in the top-right panel visualizes a sample drawn from a Gaussian signal plus a uniform background model given by eq. 5.83 and shown by the line. The remaining panels show projections of the three-dimensional posterior pdf, based on a 20,000 point MCMC chain.

we use its tools to examine, say, MLE[11] and Bayes estimators? A minimax estimator minimizes the maximum risk, and is considered one notion of a "gold-standard" desirable estimator. It can be shown for typical parametric models under weak conditions that the maximum likelihood estimator is approximately minimax, that the Bayes estimator with constant risk is also minimax, and that the MLE is approximately the Bayes estimator. Thus with enough samples each paradigm provides a good estimate, and furthermore the results are typically not very different. This is just an example of the results in the literature surrounding such analyses, which include analysis of Bayesian esimators under frequentist criteria and vice versa—see [20, 35] for some entryways into it. In the small-sample case, the prior has a larger effect, differentiating the frequentist and Bayes estimators. Whether this is good or

[11] We will sometimes use maximum likelihood as a representative of frequentist estimation, as it is the most studied, though MLE is only one option among many within frequentism.

bad leads directly to one's subjective assessment of the reasonableness of using priors, which we will come to below.

- **Uncertainty estimates**: Both MLE and Bayes estimators yield analytic confidence bands based on the asymptotic convergence of distributions to a normal distribution, for typical parametric models. For general complex models typical of machine learning, each paradigm has a way to obtain uncertainty estimates around point predictions. In Bayesian estimation, this comes in the form of posterior intervals, obtained via MCMC, and in frequentist estimation, confidence sets via resampling techniques such as the jackknife and bootstrap. Both arguably work under fairly generic circumstances, are easy to implement, and are computationally expensive. Achieving the exact guarantees of the bootstrap can require care to avoid certain model settings [6]; this is counterbalanced by the comparative difficulty in ensuring reliable results from MCMC [10], especially models with complicated likelihood landscapes. Thus in practice, neither procedure can truly be advertised as a black box that can be used in a hands-off manner.
- **Hypothesis testing**: Hypothesis testing is an area where the frequentist and Bayesian approaches differ substantially. Bayesian hypothesis testing is done by either placing a prior on the null and alternative hypotheses and computing the Bayes factor, or placing an overarching prior on both hypotheses and calculating the posterior probability of the null. Both of these rely heavily on one's confidence in the prior. In Bayesian hypothesis testing, unlike in point estimation, the influence of the prior is not attenuated with an increasing number of samples: this forces the subjective issue of the analyst's comfort level with relying on priors.
- **Goodness-of-fit testing**: Goodness-of-fit tests can be framed as hypothesis tests with no alternative hypothesis, which either makes them ill posed or useful, depending on one's point of view. Either way, they must be treated with care—if we reject the hypothesis we can conclude that we should not use the model; however if we do not reject it, we cannot conclude that the model is correct. Unlike in frequentism, in the Bayesian formulation this feature is made clear through the requirement of an alternative hypothesis. Nevertheless, the distinction is muddled by work straddling the boundaries, such as the ability to interpret frequentist goodness-of-fit tests in Bayesian terms (see [29]), the existence of Bayesian goodness-of-fit tests, and the ability to formalize goodness-of-fit tests as positing a parametric null against a nonparametric alternative; see [16, 36].
- **Nuisance parameters**: Nuisance parameters are common: we may not be interested in their values, but their values modify the distribution of our observations, and must thus be accounted for. In some cases within both frameworks, they can be eliminated analytically, and this should always be done if possible. In the Bayesian framework we can work with the joint posterior distribution using random samples drawn from it via MCMC. We can obtain the joint distribution of only the parameters of interest by simply marginalizing out the nuisance parameters. This is possible because of the availability of priors for each parameter. A frequentist approach can utilize likelihood ratio tests (see [2, 28]) to provide confidence intervals and significance tests for the parameters of interest, which account for the

presence of nuisance parameters. The Bayesian approach is more general and straightforward. On the other hand the approach relies on the quality of MCMC samples, which can be uncertain in complex models.

Subjective differences The tiebreaker, then, is to a large degree a matter of taste. These are the most common arguments for Bayesianism, over frequentism:

- **Symmetry and grand unification**: Bayesianism allows you to make probability statements about parameters. It takes the likelihood function to its logical conclusion, removes the distinction between data and parameters, thus creating a beautiful symmetry that warms the heart of many a natural scientist.
- **Extra information**: The Bayesian method gives a convenient and natural way to put in prior information. It is particularly sensible for small-sample situations where the data might be insufficient to make useful conclusions on their own. Another natural and important use arises when wanting to model the measurement error for each data point.
- **Honesty/disclosure**: The fact that we are *forced* to put in subjective priors is actually good: we always have prior beliefs, and this forces us to make them explicit. There is no such thing as an objective statistical inference procedure.
- **More elegant in practice**: Just by generating samples from the posterior, one can obtain the key results of the Bayesian method (see §5.8).

Indeed, many of the examples in this chapter illustrate the wonderful elegance possible with Bayesian methods.

These are the most common arguments against Bayesianism, for frequentism:

- **Are we really being scientific?** The posterior interval is not a true confidence interval in the sense of long-run frequencies: we cannot make statements about the true parameter with these intervals, only about our beliefs. Bayesian estimates are based on information beyond what is verifiable. Are we really going to trust, say, a major scientific conclusion based on anything but the observed data?
- **The effect of the prior is always there**: The Bayesian estimate is always biased due to the prior, even when no actual prior information (i.e., the least informative prior) is used. Bayesian hypothesis testing, in particular, is very sensitive to the choice of prior. While the effect of the prior goes away asymptotically, for any finite sample size it is still there in some unquantified way.
- **Unnecessarily complicated and computationally intensive**: The Bayesian approach requires specifying prior functions even when there is no true prior information. In practice, it often requires computationally intractable or unreliable integrals and approximations, even for models that are computationally relatively simple in the frequentist case.
- **Unnecessarily brittle and limiting**: The Bayesian method is crucially dependent on the likelihood function. As a cost function, the likelihood, while enjoying nice optimality properties when the assumed model is correct, is often highly brittle when the assumptions are wrong or there are outliers in the data. Similarly, the likelihood is often not a good choice in a nonparametric setting.

Due to the various pragmatic obstacles, it is rare for a mission-critical analysis to be done in the "fully Bayesian" manner, i.e., without the use of tried-and-true frequentist tools at the various stages. Philosophy and beauty aside, the reliability and efficiency of the underlying computations required by the Bayesian framework are the main practical issues. A central technical issue at the heart of this is that it is much easier to do optimization (reliably and efficiently) in high dimensions than it is to do integration in high dimensions. Thus the workhorse machine learning methods, while there are ongoing efforts to adapt them to Bayesian framework, are almost all rooted in frequentist methods. A work-around is to perform MAP inference, which is optimization based.

Most users of Bayesian estimation methods, in practice, are likely to use a mix of Bayesian and frequentist tools. The reverse is also true—frequentist data analysts, even if they stay formally within the frequentist framework, are often influenced by "Bayesian thinking," referring to "priors" and "posteriors." The most advisable position is probably to know both paradigms well, in order to make informed judgments about which tools to apply in which situations. Indeed, the remaining chapters in this book discuss both classical and Bayesian analysis methods. Examples of arguments for each of the respective approaches are [4] and [22], and a more recent attempt at synthesis can be found in [5]. For more about the relative merits of the frequentist and Bayesian frameworks, please see Greg05 and Wass10.

References

[1] Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician 39*(2), 83–87.

[2] Cash, W. (1979). Parameter estimation in astronomy through application of the likelihood ratio. *ApJ 228*, 939–947.

[3] Clyde, M. A., J. O. Berger, F. Bullard, and others (2007). Current challenges in Bayesian model choice. In G. J. Babu and E. D. Feigelson (Eds.), *Statistical Challenges in Modern Astronomy IV*, Volume 371 of *Astronomical Society of the Pacific Conference Series*, pp. 224.

[4] Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician 40*(1), 1–5.

[5] Efron, B. (2004). Bayesians, frequentists, and scientists. Text of the 164th ASA presidential address, delivered at the awards ceremony in Toronto on August 10.

[6] Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

[7] Foreman-Mackey, D., D. W. Hogg, D. Lang, and J. Goodman (2012). emcee: The MCMC Hammer. *ArXiv:astro-ph/1202.3665*.

[8] Franklin, J. (2001). *The Science of Conjecture: Evidence and Probability Before Pascal*. Johns Hopkins University Press.

[9] Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian Data Analysis*. Chapman and Hall.

[10] Gilks, W., S. Richardson, and D. Spiegelhalter (1995). *Markov Chain Monte Carlo in Practice*. Chapman and
Hall/CRC Interdisciplinary Statistics. Chapman and Hall/CRC.

[11] Goodman, J. and J. Weare (2010). Ensemble samplers with affine invariance. *Commun. Appl. Math. Comput. Sci. 5*, 65–80.

[12] Gregory, P. C. (2007). A Bayesian Kepler periodogram detects a second planet in HD208487. *MNRAS 374*, 1321–1333.
[13] Gregory, P. C. and T. J. Loredo (1992). A new method for the detection of a periodic signal of unknown shape and period. *ApJ 398*, 146–168.
[14] Hogg, D. W. (2012). Data analysis recipes: Probability calculus for inference. *ArXiv:astro-ph/1205.4446*.
[15] Ivezić, Ž., J. A. Tyson, E. Acosta, and others (2008). LSST: From science drivers to reference design and anticipated data products. *ArXiv:astro-ph/0805.2366*.
[16] Johnson, V. E. (2004). A Bayesian chi-squared test for goodness-of-fit. *Annals of Statistics 32*(6), 2361–2384.
[17] Kelly, B. C., R. Shetty, A. M. Stutz, and others (2012). Dust spectral energy distributions in the era of Herschel and Planck: A hierarchical Bayesian-fitting technique. *ApJ 752*, 55.
[18] Kessler, R., A. C. Becker, D. Cinabro, and others (2009). First-year Sloan Digital Sky Survey-II supernova results: Hubble diagram and cosmological parameters. *ApJS 185*, 32–84.
[19] Knuth, K. H. (2006). Optimal data-based binning for histograms. *ArXiv:physics/0605197*.
[20] Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2nd ed.). Springer Texts in Statistics. Springer.
[21] Liddle, A. R. (2007). Information criteria for astrophysical model selection. *MNRAS 377*, L74–L78.
[22] Loredo, T. J. (1992). Promise of Bayesian inference for astrophysics. In E. D. Feigelson and G. J. Babu (Eds.), *Statistical Challenges in Modern Astronomy*, pp. 275–306.
[23] Loredo, T. J. (1999). Computational technology for Bayesian inference. In D. M. Mehringer, R. L. Plante, and D. A. Roberts (Eds.), *Astronomical Data Analysis Software and Systems VIII*, Volume 172 of *Astronomical Society of the Pacific Conference Series*, pp. 297.
[24] Lutz, T. E. and D. H. Kelker (1973). On the use of trigonometric parallaxes for the calibration of luminosity systems: Theory. *PASP 85*, 573.
[25] Marquis de Laplace (1995). *A Philosophical Essay on Probabilities*. Dover.
[26] McGrayne, S. B. (2011). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press.
[27] Press, W. H. (1997). Understanding data better with Bayesian and global statistical methods. In J. N. Bahcall, and J. P. Ostriker (Ed.), *Unsolved Problems in Astrophysics*. Princeton University Press.
[28] Protassov, R., D. A. van Dyk, A. Connors, V. L. Kashyap, and A. Siemiginowska (2002). Statistics, handle with care: Detecting multiple model components with the likelihood ratio test. *ApJ 571*, 545–559.
[29] Rubin, H. and J. Sethuraman (1965). Bayes risk efficiency. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002) 27*, 347–356.
[30] Scargle, J. D. (1998). Studies in astronomical time series analysis. V. Bayesian blocks, a new method to analyze structure in photon counting data. *ApJ 504*, 405.
[31] Scargle, J. D., J. P. Norris, B. Jackson, and J. Chiang (2012). Studies in astronomical time series analysis. VI. Bayesian block representations. *ArXiv:astro-ph/1207.5578*.
[32] Smith, H. (2003). Is there really a Lutz-Kelker bias? Reconsidering calibration with trigonometric parallaxes. *MNRAS 338*, 891–902.

[33] Smith, Jr., H. (1987). The calibration problem. I - Estimation of mean absolute magnitude using trigonometric parallaxes. II - Trigonometric parallaxes selected according to proper motion and the problem of statistical parallaxes. *A&A 171*, 336–347.

[34] Teerikorpi, P. (1997). Observational selection bias affecting the determination of the extragalactic distance scale. *ARAA 35*, 101–136.

[35] van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[36] Verdinelli, I. and L. Wasserman (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Annals of Statistics 26*, 1215–1241.